*Biometrika* (2020), **00**, 0, *pp.* 1–15
*Printed in Great Britain*
doi: 10.1093/biomet/asaa044



# Large-sample asymptotics of the pseudo-marginal method

BY S. M. SCHMON, G. DELIGIANNIDIS, A. DOUCET
*Department of Statistics, University of Oxford, 24–29 St Giles', Oxford OX1 3LB, U.K.*
schmon@stats.ox.ac.uk deligiannidis@stats.ox.ac.uk doucet@stats.ox.ac.uk

AND M. K. PITT
*Department of Mathematics, King's College London, Strand, London WC2R 2LS, U.K.*
michael.pitt@kcl.ac.uk

SUMMARY

The pseudo-marginal algorithm is a variant of the Metropolis–Hastings algorithm which samples asymptotically from a probability distribution when it is only possible to estimate unbiasedly an unnormalized version of its density. Practically, one has to trade off the computational resources used to obtain this estimator against the asymptotic variances of the ergodic averages obtained by the pseudo-marginal algorithm. Recent works on optimizing this trade-off rely on some strong assumptions, which can cast doubts over their practical relevance. In particular, they all assume that the distribution of the difference between the log-density, and its estimate is independent of the parameter value at which it is evaluated. Under regularity conditions we show that as the number of data points tends to infinity, a space-rescaled version of the pseudo-marginal chain converges weakly to another pseudo-marginal chain for which this assumption indeed holds. A study of this limiting chain allows us to provide parameter dimension-dependent guidelines on how to optimally scale a normal random walk proposal, and the number of Monte Carlo samples for the pseudo-marginal method in the large-sample regime. These findings complement and validate currently available results.

*Some key words*: Asymptotic posterior normality; Intractable likelihood; Large-sample theory; Metropolis–Hastings algorithm; Random measure; Weak convergence.

## 1. INTRODUCTION

The pseudo-marginal algorithm is a variant of the popular Metropolis–Hastings algorithm in which an unnormalized version of the target density is replaced with a nonnegative unbiased estimate. The algorithm first appeared in the physics literature (Lin et al., 2000) and has become popular in Bayesian statistics, as many intractable likelihood functions can be estimated unbiasedly using importance sampling or particle filters (Beaumont, 2003; Andrieu & Roberts, 2009; Andrieu et al., 2010).

Replacing the true likelihood in the Metropolis–Hastings algorithm with an estimate results in a trade-off: the asymptotic variance of an ergodic average of a pseudo-marginal chain typically decreases as the number of Monte Carlo samples, $N$, used to obtain the likelihood estimator increases, as established by Andrieu & Vihola (2016) for importance sampling estimators; however, this comes at the cost of a higher computational burden. An important task in practice is therefore to choose $N$ such that the computational resources required to obtain a given asymptotic variance are minimized. This problem has already been investigated by Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015), who obtained guidelines under various assumptions either on the proposal (Pitt et al., 2012; Doucet et al., 2015) or on the proposal and target distribution (Sherlock et al., 2015).

Additionally, all these contributions make the assumption that the noise in the loglikelihood estimator, i.e., the difference between the estimator and the true loglikelihood, is Gaussian with variance inversely proportional to $N$, its mean and variance being independent of the parameter value at which it is evaluated. A similar assumption has also been used by Nemeth et al. (2016) for the analysis of a related algorithm. This assumption can cast doubts over the practical relevance of the guidelines provided in these works. The normal-noise assumption was motivated in Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015) by the fact that the error in the loglikelihood estimator for state-space models computed using a particle filter is asymptotically normal with variance proportional to $\gamma$ as $T \to \infty$ with $N = T/\gamma$ (Bérard et al., 2014). In addition, the constant-variance assumption over the parameter space was motivated in Pitt et al. (2012) and Doucet et al. (2015) by the fact that the posterior typically concentrates as $T$ increases. However, no formal argument justifying why the pseudo-marginal chain would behave as a Markov chain for which these assumptions hold has been provided.

In this article we carry out a novel weak convergence analysis of the pseudo-marginal algorithm in a Bayesian setting, which not only justifies these assumptions, but also allows us to obtain more precise guidelines on how to optimally tune the algorithm as a function of the parameter dimension $d$. Weak convergence techniques have become very popular in the Markov chain Monte Carlo literature since their introduction in the seminal paper of Roberts et al. (1997). With the exception of Deligiannidis et al. (2018), all these analyses have been performed in the asymptotic regime, where the parameter dimension $d$ tends to infinity. Results of this type typically require strong structural assumptions on the target distribution, such as having $d$ independent and identically distributed components as in Sherlock et al. (2015). We analyse here the pseudo-marginal scheme in the large-sample asymptotic regime where the number of data points $T$ goes to infinity while $d$ is fixed. Under weak regularity conditions, we show that a space-rescaled version of the pseudo-marginal chain converges to a pseudo-marginal chain targeting a normal distribution for which the noise in the loglikelihood estimator is indeed also normal with constant mean and variance. We provide numerical results on optimally scaling normal random walk proposals and the noise variance to optimize the performance of the limiting Markov chain as a function of $d$. These guidelines complement and validate the results obtained in Doucet et al. (2015) and Sherlock et al. (2015). All proofs can be found in the Supplementary Material.

## 2. The pseudo-marginal algorithm

### 2.1. *Background*

Consider a Bayesian model on the Borel space $\{\Theta, \mathcal{B}(\Theta)\}$ where $\Theta \subseteq \mathbb{R}^d$. The parameter $\theta \in \Theta$ follows a prior distribution $p(\mathrm{d}\theta)$, and $\theta \mapsto p(y \mid \theta)$ is the likelihood function, where $y = (y_1, \ldots, y_T)$ denotes the vector of observations. When the likelihood arises from a complex latent variable model, an analytic expression for $p(y \mid \theta)$ may not be available. Hence, the standard Metropolis–Hastings algorithm cannot be used to sample the posterior distribution $p(\mathrm{d}\theta \mid y) \propto p(\mathrm{d}\theta)\, p(y \mid \theta)$, as the likelihood ratio $p(y \mid \theta')/p(y \mid \theta)$ appearing in the Metropolis–Hastings acceptance probability, when at parameter $\theta$ and proposing $\theta'$, cannot be computed. Assume that we have access to an unbiased positive estimator $\hat{p}(y \mid \theta, U)$ of the intractable likelihood $p(y \mid \theta)$, where $U \sim m_\theta$ represents the auxiliary variables on $\{\mathcal{U}, \mathcal{B}(\mathcal{U})\}$ used to compute this estimator. We introduce the following probability measure on $\{\Theta\times\mathcal{U},\ \mathcal{B}(\Theta)\times\mathcal{B}(\mathcal{U})\}$:

$$\pi(\mathrm{d}\theta,\mathrm{d}u) = p(\mathrm{d}\theta \mid y)\frac{\hat{p}(y \mid \theta, u)}{p(y \mid \theta)}\, m_\theta(\mathrm{d}u),$$

which satisfies $\pi(\mathrm{d}\theta) = p(\mathrm{d}\theta \mid y)$. The pseudo-marginal algorithm is a Metropolis–Hastings scheme targeting $\pi(\mathrm{d}\theta,\mathrm{d}u)$, and hence marginally $p(\mathrm{d}\theta \mid y)$, using a proposal distribution $Q(\theta,u;\mathrm{d}\theta',\mathrm{d}u') = q(\theta,\mathrm{d}\theta')\, m_{\theta'}(\mathrm{d}u')$. This yields the acceptance probability

$$\alpha(\theta,u;\theta',u') = \min\left\{1,\ r(\theta,\theta')\frac{\hat{p}(y \mid \theta',u')/p(y \mid \theta')}{\hat{p}(y \mid \theta,u)/p(y \mid \theta)}\right\}, \quad r(\theta,\theta') = \frac{\pi(\mathrm{d}\theta')}{\pi(\mathrm{d}\theta)}\frac{q(\theta',\mathrm{d}\theta)}{q(\theta,\mathrm{d}\theta')}.$$

As in previous works (Andrieu & Roberts, 2009; Pitt et al., 2012; Andrieu & Vihola, 2015; Doucet et al., 2015; Sherlock et al., 2015), we analyse the pseudo-marginal algorithm using additive noise in the loglikelihood estimator, writing $Z(\theta) = \log \hat{p}(y \mid \theta, U) - \log p(y \mid \theta)$. This parameterization allows us to write the target distribution as a measure on $\{\Theta \times \mathbb{R},\ \mathcal{B}(\Theta)\times\mathcal{B}(\mathbb{R})\}$ with

$$\pi(\mathrm{d}\theta,\mathrm{d}z) = p(\mathrm{d}\theta \mid y)\exp(z)\, g(\mathrm{d}z \mid \theta),$$

where $Z(\theta) \sim g(\cdot \mid \theta)$ when $U \sim m_\theta$, and the associated pseudo-marginal kernel is

$$P(\theta,z;\mathrm{d}\theta',\mathrm{d}z') = q(\theta,\mathrm{d}\theta')\, g(\mathrm{d}z' \mid \theta')\, \alpha\big(\theta,z;\theta',z'\big) + \rho(\theta,\ z)\,\delta_{(\theta,\ z)}(\mathrm{d}\theta',\mathrm{d}z'),$$

with acceptance probability

$$\alpha\big(\theta,z;\theta',z'\big) = \min\big\{1,\ r(\theta,\theta')\exp(z'-z)\big\}$$

and corresponding rejection probability $\rho(\theta,\ z)$.

## 2.2. *Literature review*

In this subsection we review recent research motivating the present work. To this end, we need to introduce some additional notation. Let $\mu$ be a probability measure on $\{\mathbb{R}^n,\mathcal{B}(\mathbb{R}^n)\}$ and $\Pi\colon \mathbb{R}^n \times \mathcal{B}(\mathbb{R}^n) \to [0,1]$ a Markov transition kernel. For any measurable function $f$ and measurable set $A$, we write $\mu(f) = \int f(x)\mu(\mathrm{d}x)$, $\mu(A) = \mu\{\mathbb{I}_A(\cdot)\}$ and $\Pi f(x) = \int \Pi(x,\mathrm{d}y)f(y)$. We consider the Hilbert space $L^2(\mu)$ with inner product $\langle f,g\rangle_\mu = \int f(x)g(x)\mu(\mathrm{d}x)$. For a function $f \in L^2(\mu)$, the asymptotic variance of averages of a stationary Markov chain $(X_k)_{k\geqslant 0}$ of a $\mu$-invariant transition kernel $\Pi$ is defined as

$$\mathrm{var}(f,\Pi) = \lim_{M\to\infty}\frac{1}{M}E\left[\left\{\sum_{k=1}^{M} f(X_k) - \mu(f)\right\}^2\right],$$

or $\mathrm{var}(f,\Pi) = \mathrm{var}\{f(X_0)\}\,\textsc{iat}(f,\Pi)$ when the integrated autocorrelation time

$$\text{IAT}(f, \Pi) = 1 + 2 \sum_{k=1}^{\infty} \frac{\text{cov}\{f(X_0), f(X_k)\}}{\text{var}\{f(X_0)\}}$$

is finite. We denote by $\varphi(x; m, \Lambda)$ the normal density of argument $x$, mean $m$ and covariance $\Lambda$.

To obtain guidelines for balancing the computational cost and accuracy of the likelihood estimator, Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015) made the simplifying assumption that $g(\mathrm{d}z \mid \theta) = \varphi(\mathrm{d}z; -\sigma^2/2, \sigma^2)$ with $\sigma^2 \propto 1/N$ and focused on functions $f \in L^2(\pi)$ such that $f(\theta, z) = f(\theta, z')$ for any $z$ and $z'$. Under these assumptions, it was first proposed by Pitt et al. (2012) to minimize

$$\text{CT}(f, P_\sigma) = \frac{\text{IAT}(f, P_\sigma)}{\sigma^2} \tag{1}$$

with respect to $\sigma$, where

$$P_\sigma(\theta, z; \mathrm{d}\theta', \mathrm{d}z') = q(\theta, \mathrm{d}\theta')\, \varphi(\mathrm{d}z; -\sigma^2/2, \sigma^2)\, \alpha\big(\theta, z; \theta', z'\big) + \rho_\sigma(\theta, z)\, \delta_{(\theta, z)}(\mathrm{d}\theta', \mathrm{d}z'), \tag{2}$$

with $\rho_\sigma(\theta, z)$ being the corresponding rejection probability. Criterion (1) arises from the fact that the computational time required to evaluate the likelihood is typically proportional to $N$. Under the additional assumption that $q(\theta, \mathrm{d}\theta') = \pi(\mathrm{d}\theta')$, the minimizer of $\text{CT}(f, P_\sigma)$ is $\sigma = 0.92$ (Pitt et al., 2012). For general proposal distributions, Doucet et al. (2015) minimize upper bounds on $\text{CT}(f, P_\sigma)$. This results in guidelines saying that one should indeed select $\sigma$ around 1.0 when the Metropolis–Hastings algorithm using the exact likelihood would provide an estimator having a small integrated autocorrelation time, and around 1.7 when this autocorrelation time is very large (Doucet et al., 2015). In practical scenarios, the integrated autocorrelation time of the Metropolis–Hastings algorithm using the exact likelihood is unknown, and the results in Doucet et al. (2015) suggest selecting $\sigma$ around 1.2 as a robust default choice. A slightly different approach was taken by Sherlock et al. (2015). In addition to similar noise assumptions, they assumed that the posterior factorizes into $d$ independent and identically distributed components and that one uses an isotropic normal random walk proposal of jump size proportional to $\ell$. In this context, one maximizes with respect to $(\sigma, \ell)$ the expected squared jump distance associated with the pseudo-marginal sequence of the first parameter component $(\vartheta_{1,k})_{k\geqslant 0}$ divided by the noise variance as $d \to \infty$. In this asymptotic regime, a time-rescaled version of $(\vartheta_{1,k})_{k\geqslant 0}$ converges weakly to a diffusion process and the adequately rescaled expected squared jump distance converges to the squared diffusion coefficient of this process. Maximizing this squared jump distance is asymptotically equivalent to minimizing $\text{CT}(f, P_\sigma)$ irrespective of $f$ (see Roberts & Rosenthal, 2014), and its maximizing arguments are $\sigma = 1.8$ and $\ell = 2.56$ (Sherlock et al., 2015, Corollary 1). In practice, the standard deviation of the loglikelihood estimator varies over the parameter space and one selects $N$ such that this standard deviation is approximately equal to the desired $\sigma$ for a parameter value around the mode of the posterior obtained from a preliminary run.

The strong assumptions made in those papers can bring into question the merits of the guidelines they provide. Our novel weak convergence analysis of the pseudo-marginal algorithm justifies the main common assumption in the large-sample regime, as $T \to \infty$. This convergence occurs under fairly weak regularity conditions on the posterior distribution. The resulting limiting algorithms can be optimized to give guidelines for random walk proposals without relying on any upper bound as in Doucet et al. (2015).

## 3. Large-sample asymptotics of the pseudo-marginal algorithm

### 3.1. *Notation and assumptions*

Our analysis of the pseudo-marginal algorithm relies on the assumption that the posterior concentrates, which is most commonly formulated in terms of convergence in probability with respect to the data distribution, denoted by $\mathbb{P}^Y$. For our result to hold under this weak assumption, we take into account the randomness induced by the data, resulting in a random Markov chain and requiring us to deal with weak convergence of random probability measures. To make this more precise, we introduce the following notation.

The observations $(Y_t)_{t\geqslant 1}$ are regarded as random variables defined on a probability space $\{\mathsf{Y}^{\mathbb{N}}, \mathcal{B}(\mathsf{Y})^{\mathbb{N}}, \mathbb{P}^Y\}$, where $\mathcal{B}(\mathsf{Y})^{\mathbb{N}}$ denotes the Borel $\sigma$-algebra and we write $\Omega = \mathsf{Y}^{\mathbb{N}}$ for brevity. For $T \geqslant 1$ we can define the random variables $Y_{1:T} = (Y_1, \ldots, Y_T)$ as the coordinate projections to $\mathsf{Y}^T$. Then, for $\omega = (y_t)_{t\geqslant 1} \in \Omega$, $\pi_T^\omega(\mathrm{d}\theta) = p(\mathrm{d}\theta \mid y_{1:T})$ denotes a regular version of the target posterior distribution and, for any $\theta \in \Theta$, $g_T^\omega(\mathrm{d}z \mid \theta)$ denotes the conditional distribution of the error in the loglikelihood estimator given observations $y_{1:T}$. The measures $\pi_T^\omega$ and $g_T^\omega$ can be interpreted as random measures. Relevant results for random measures are discussed briefly in § 4 and in more detail in the Supplementary Material. In the following we will use a superscript $\omega$ to highlight that a certain quantity depends on the data. All probability densities considered hereafter are with respect to Lebesgue measure, and we use the same symbols for distributions and densities; for example, $\mu(\mathrm{d}\theta) = \mu(\theta)\,\mathrm{d}\theta$.

In this context, the target distribution of the pseudo-marginal algorithm is

$$\pi_T^\omega(\mathrm{d}\theta, \mathrm{d}z) = \pi_T^\omega(\mathrm{d}\theta)\exp(z)\, g_T^\omega(\mathrm{d}z \mid \theta),$$

and its transition kernel is

$$P_T^\omega(\theta, z; \mathrm{d}\theta', \mathrm{d}z') = q_T(\theta, \mathrm{d}\theta')\, g_T^\omega(\mathrm{d}z' \mid \theta')\, \alpha_T^\omega\big(\theta, z; \theta', z'\big) + \rho_T^\omega(\theta,\ z)\, \delta_{(\theta,\ z)}(\mathrm{d}\theta', \mathrm{d}z'),$$

where

$$\alpha_T^\omega\big(\theta, z; \theta', z'\big) = \min\left\{1,\ \frac{\pi_T^\omega(\mathrm{d}\theta')}{\pi_T^\omega(\mathrm{d}\theta)}\frac{q_T(\theta', \mathrm{d}\theta)}{q_T(\theta, \mathrm{d}\theta')}\exp(z' - z)\right\}$$

and $\rho_T^\omega(\theta,\ z)$ is the corresponding rejection probability.

Our first assumption is that the posterior distributions concentrate towards a Gaussian at rate $1/\sqrt{T}$. We denote by $\mathcal{Y}_T$ the $\sigma$-algebra spanned by $Y_{1:T}$.

*Assumption* 1. The posterior distributions $\{\pi_T^\omega(\mathrm{d}\theta)\}_{T\geqslant 1}$ admit Lebesgue densities, and there exist a $d \times d$ positive-definite matrix $\Sigma$, a parameter value $\bar\theta \in \Theta$ and a sequence $(\hat\theta_T^\omega)_{T\geqslant 1}$ of $\mathcal{Y}_T$-adapted random variables such that as $T \to \infty$,

$$\int \left|\pi_T^\omega(\theta) - \varphi\big(\theta; \hat\theta_T^\omega, \Sigma/T\big)\right| \mathrm{d}\theta \to 0, \quad \hat\theta_T^\omega \to \bar\theta, \tag{3}$$

both limits being in $\mathbb{P}^Y$-probability.

Assumption 1 is satisfied if a Bernstein–von Mises theorem holds; see van der Vaart (2000, Theorem 10.1) and Kleijn & Van der Vaart (2012). Our second assumption is that we use random walk proposal distributions with appropriately scaled increments.

*Assumption* 2. The proposal distributions $\{q_T(\theta, \mathrm{d}\theta')\}_{T\geqslant 1}$ admit densities of the form

$$q_T(\theta,\theta') = \surd T \nu\big\{\surd T(\theta'-\theta)\big\},$$

where $\nu$ is a continuous density on $\mathbb{R}^d$.

Finally, we assume that the error in the loglikelihood estimator satisfies a central limit theorem conditional on $\mathcal{Y}_T$ and that this convergence holds uniformly in a neighbourhood of $\bar\theta$.

*Assumption* 3. There exists an $\varepsilon$-ball $B(\bar\theta)$ around $\bar\theta$ such that the distributions of the error in the loglikelihood estimator $\{g_T^\omega(\mathrm{d}z \mid \theta)\}_{T\geqslant 1}$ satisfy, as $T\to\infty$,

$$\sup_{\theta\in B(\bar\theta)} d_{\mathrm{BL}}\big[g_T^\omega(\,\cdot\mid\theta),\ \varphi\{\,\cdot\,; -\sigma^2(\theta)/2, \sigma^2(\theta)\}\big] \to 0$$

in $\mathbb{P}^Y$-probability, where $d_{\mathrm{BL}}(\cdot\,,\cdot)$ denotes the bounded Lipschitz metric and the function $\sigma : \Theta\to[0,\infty)$ is continuous at $\bar\theta$ with $0<\sigma(\bar\theta)<\infty$. An analogous result holds for $\bar g_T^\omega(\mathrm{d}z\mid\theta) = \exp(z) g_T^\omega(\mathrm{d}z\mid\theta)$, the distribution of this error at equilibrium; that is, as $T\to\infty$,

$$\sup_{\theta\in B(\bar\theta)} d_{\mathrm{BL}}\big[\bar g_T^\omega(\,\cdot\mid\theta),\ \varphi\{\,\cdot\,; \sigma^2(\theta)/2, \sigma^2(\theta)\}\big] \to 0$$

in $\mathbb{P}^Y$-probability.

We will refer to convergence in probability with respect to the bounded Lipschitz metric as weak convergence in probability. In § 5 we provide sufficient conditions under which Assumption 3 is satisfied for random effects models, where the likelihood estimator is a product of $T$ independent importance sampling estimators. This differs from scenarios in which the likelihood estimator is given by one single importance sampling estimator, as in Sherlock et al. (2017). Empirical evidence in Pitt et al. (2012) and Doucet et al. (2015) also suggests that Assumption 3 might hold for a large class of state-space models when the likelihood is estimated using particle filters. Under strong assumptions, a standard central limit theorem has been established by Bérard et al. (2014) for $g_T^\omega(\,\cdot\mid\theta)$. However, it would be technically very challenging to provide weak sufficient conditions under which Assumption 3 holds in this context.

### 3.2. *Weak convergence in the large-sample regime*

Denote by $(\vartheta_{T,k}^\omega, Z_{T,k}^\omega)_{k\geqslant 0}$ the stationary Markov chain defined by the pseudo-marginal kernel, $(\vartheta_{T,0}^\omega, Z_{T,0}^\omega)\sim\pi_T^\omega$ and $(\vartheta_{T,k}^\omega, Z_{T,k}^\omega)\sim P_T^\omega(\vartheta_{T,k-1}^\omega, Z_{T,k-1}^\omega;\cdot)$ for $k\geqslant 1$. Let $\chi_T^\omega = (\tilde\vartheta_{T,k}^\omega, Z_{T,k}^\omega)_{k\geqslant 0}$ where $\tilde\vartheta_{T,k}^\omega = \surd T(\vartheta_{T,k}^\omega - \hat\theta_T^\omega)$ is the Markov chain arising from rescaling the parameter component of the pseudo-marginal chain. Its transition kernel is thus

$$\tilde P_T^\omega(\tilde\theta, z; \mathrm{d}\tilde\theta', \mathrm{d}z') = \tilde q_T(\tilde\theta, \mathrm{d}\tilde\theta')\, \tilde g_T^\omega(\mathrm{d}z'|\tilde\theta')\, \tilde\alpha_T^\omega\big(\tilde\theta, z; \tilde\theta', z'\big) + \tilde\rho_T^\omega(\tilde\theta, z)\, \delta_{(\tilde\theta,z)}(\mathrm{d}\tilde\theta', \mathrm{d}z'), \quad (4)$$

where

$$\tilde\alpha_T^\omega\big(\tilde\theta, z; \tilde\theta', z'\big) = \min\left\{1,\ \frac{\tilde\pi_T^\omega(\mathrm{d}\tilde\theta')}{\tilde\pi_T^\omega(\mathrm{d}\tilde\theta)}\frac{\tilde q_T(\tilde\theta', \mathrm{d}\tilde\theta)}{\tilde q_T(\tilde\theta, \mathrm{d}\tilde\theta')}\exp(z'-z)\right\},$$

$\tilde{\rho}_T^{\omega}(\theta,\, z)$ is the corresponding rejection probability, $\tilde{\pi}_T^{\omega}(\tilde{\theta}) = \pi_T^{\omega}(\hat{\theta}_T^{\omega} + \tilde{\theta}/\sqrt{T})/\sqrt{T}$, $\tilde{q}_T(\tilde{\theta},\tilde{\theta}') = q_T(\hat{\theta}_T^{\omega} + \tilde{\theta}/\sqrt{T}, \hat{\theta}_T^{\omega} + \tilde{\theta}'/\sqrt{T})/\sqrt{T}$ and $\tilde{g}_T^{\omega}(z \mid \tilde{\theta}) = g_T^{\omega}(z \mid \hat{\theta}_T^{\omega} + \tilde{\theta}/\sqrt{T})$. Under Assumption 2 we have $\tilde{q}_T(\tilde{\theta},\tilde{\theta}') = \nu(\tilde{\theta}' - \tilde{\theta}) = \tilde{q}(\tilde{\theta},\tilde{\theta}')$. We now state the main result of this paper.

Theorem 1. *Under Assumptions 1–3, as $T \to \infty$ the sequence $(\chi_T^{\omega})_{T\geqslant 1}$ of stationary Markov chains converges weakly in $\mathbb{P}^Y$-probability to the law of a stationary Markov chain with initial distribution*

$$\tilde{\pi}(\mathrm{d}\tilde{\theta}, \mathrm{d}z) = \varphi(\mathrm{d}\tilde{\theta}; 0, \Sigma)\, \varphi(\mathrm{d}z; \sigma^2/2, \sigma^2) \tag{5}$$

*and transition kernel*

$$\tilde{P}(\tilde{\theta}, z; \mathrm{d}\tilde{\theta}', \mathrm{d}z') = \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}')\, \varphi(\mathrm{d}z'; -\sigma^2/2, \sigma^2)\, \tilde{\alpha}(\tilde{\theta}, z; \tilde{\theta}', z') + \tilde{\rho}(\tilde{\theta}, z)\, \delta_{(\tilde{\theta},z)}(\mathrm{d}\tilde{\theta}', \mathrm{d}z'), \tag{6}$$

*where $\sigma = \sigma(\bar{\theta})$,*

$$\tilde{\alpha}(\tilde{\theta}, z; \tilde{\theta}', z') = \min\left\{1,\ \frac{\varphi(\tilde{\theta}'; 0, \Sigma)}{\varphi(\tilde{\theta}; 0, \Sigma)} \frac{\tilde{q}(\tilde{\theta}', \tilde{\theta})}{\tilde{q}(\tilde{\theta}, \tilde{\theta}')} \exp(z' - z)\right\}$$

*and $\tilde{\rho}(\theta,\, z)$ is the corresponding rejection probability.*

Under this asymptotic regime, the limiting transition kernel $\tilde{P}$ in (6) is also a pseudo-marginal kernel for which the noise distribution is $\varphi(\mathrm{d}z; -\sigma^2/2, \sigma^2)$ as assumed in previous analyses (Pitt et al., 2012; Doucet et al., 2015; Sherlock et al., 2015). As Theorem 1 is a weak convergence result, it does not imply that the integrated autocorrelation time of the pseudo-marginal kernel $\tilde{P}_T^{\omega}$ converges to that of $\tilde{P}$. However, it suggests that for large $T$, some characteristics of $\tilde{P}_T^{\omega}$ can indeed be captured by those of the kernel (2), which can be obtained from $\tilde{P}$ by using the change of variables $\theta = \hat{\theta}_T^{\omega} + \tilde{\theta}/\sqrt{T}$ and substituting the true target for its normal approximation $\varphi(\theta; \hat{\theta}_T^{\omega}, \Sigma/T)$, hence removing a level of approximation.

## 4. Outline of the proof of the main result

### 4.1. *Random Markov chains*

The proof of Theorem 1 follows from a slightly more general result on weak convergence of random Markov chains on Polish spaces, given in Theorem 2 below. We first introduce some notation and recall some definitions concerning random probability measures needed to define random Markov chains; see the Supplementary Material or Crauel (2003) for more details.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $S$ a Polish space endowed with its Borel $\sigma$-algebra $\mathcal{B}(S)$. We equip the product space $\Omega \times S$ with the product $\sigma$-algebra $\mathcal{F} \otimes \mathcal{B}(S)$. We denote by $\mathcal{P}(S)$ the space of Borel probability measures which is itself endowed with the Borel $\sigma$-algebra $\mathcal{B}\{\mathcal{P}(S)\}$ generated by the weak topology. Finally, $C_{\mathrm{b}}(S)$ denotes the set of continuous bounded functions and $\mathrm{BL}(S)$ the set of bounded Lipschitz functions.

Definition 1. *A random probability measure is a map $\mu\colon \Omega \times \mathcal{B}(S) \to [0, 1]$, $(\omega, B) \mapsto \mu(\omega, B) = \mu^{\omega}(B)$, such that for every $B \in \mathcal{B}(S)$ the map $\omega \mapsto \mu(\omega, B)$ is measurable while $\mu^{\omega} \in \mathcal{P}(S)$ $\mathbb{P}$-almost surely.*

For all bounded and measurable functions $g\colon \Omega \times S \to \mathbb{R}$, $\omega \mapsto \int_S g(\omega, x)\mu^{\omega}(\mathrm{d}x)$ is measurable (Crauel, 2003, Proposition 3.3) and hence the map $\omega \mapsto \mu^{\omega}(f)$ is a random variable for bounded measurable functions $f\colon S \to \mathbb{R}$. Consequently, $\mu^{\omega}\colon \Omega \to \mathcal{P}(S)$ is a Borel-measurable map. Conversely, it can be shown that any random element of $\{\mathcal{P}(S), \mathcal{B}\{\mathcal{P}(S)\}\}$ fulfils the conditions set out in Definition 1; see Crauel (2003, Remark 3.20(i)) or Kallenberg (2006, Lemma 1.37).

DEFINITION 2. *A random Markov kernel is a map* $K\colon \Omega \times S \times \mathcal{B}(S) \to [0,1]$, $(\omega, x, B) \mapsto K(\omega, x, B) = K^{\omega}(x, B)$, *such that*

(i) $(\omega, x) \mapsto K^{\omega}(x, B)$ *is* $\mathcal{F} \otimes \mathcal{B}(S)$*-measurable for every* $B \in \mathcal{B}(S)$, *and*
(ii) $K^{\omega}(x, \cdot) \in \mathcal{P}(S)$ $\mathbb{P}$*-almost surely for every* $x \in S$.

LEMMA 1. *Given a random probability measure* $\mu^{\omega}$ *and random Markov kernel* $K^{\omega}$, *there exists an almost surely unique random probability measure* $\mu^{\mathbb{N},\omega}$ *on* $S^{\mathbb{N}}$ *such that*

$$\mu^{\mathbb{N},\omega}(A_1 \times \cdots \times A_k \times E_{k+1}) = \int_{A_1} \mu^{\omega}(\mathrm{d}x_1) \int_{A_2} K^{\omega}(x_1, \mathrm{d}x_2) \cdots \int_{A_k} K^{\omega}(x_{k-1}, \mathrm{d}x_k)$$

*for any* $A_i \in \mathcal{B}(S)$ $(i = 1, \ldots, k)$, $k \in \mathbb{N}$ *and* $E_{k+1} = \times_{i=k+1}^{\infty} S$.

### 4.2. *Convergence of random Markov chains*

For a sequence of random probability measures $(\mu_n^{\omega})_{n\geqslant 1}$ converging in a suitable sense to a probability measure $\mu$, and a sequence of random Markov kernels $(K_n^{\omega})_{n\geqslant 1}$ converging in a suitable sense to a Markov kernel $K$, we show that the distributions of the associated Markov chains $(\mu_n^{\mathbb{N},\omega})_{n\geqslant 1}$ defined in Lemma 1 converge weakly in probability to the distribution $\mu^{\mathbb{N}}$ of the homogeneous Markov chain of the initial distribution $\mu$ and Markov kernel $K$.

THEOREM 2. *Suppose that the following assumptions hold:*

(i) *the random probability measures* $(\mu_n^{\omega})_{n\geqslant 1}$ *converge weakly in probability to a probability measure* $\mu$ *as* $n \to \infty$;
(ii) *the random Markov transition kernels* $(K_n^{\omega})_{n\geqslant 1}$ *satisfy*

$$\int \left| K_n^{\omega} f(x) - Kf(x) \right| \mu_n^{\omega}(\mathrm{d}x) \to 0$$

*in probability as* $n \to \infty$ *for all* $f \in \mathrm{BL}(S)$, *where* $K$ *is a Markov transition kernel;*
(iii) *the transition kernel* $K$ *is such that* $x \mapsto Kf(x)$ *is continuous for any* $f \in C_{\mathrm{b}}(S)$.

*Then, as* $n \to \infty$, *the measures* $(\mu_n^{\mathbb{N},\omega})_{n\geqslant 1}$ *on* $S^{\mathbb{N}}$ *converge weakly in probability to the measure* $\mu^{\mathbb{N}}$ *induced by the Markov chain with initial distribution* $\mu$ *and transition kernel* $K$.

### 4.3. *Application to the pseudo-marginal algorithm*

Theorem 1 follows from Theorem 2 upon verifying that, under Assumptions 1–3, all the conditions set out in Theorem 2 are fulfilled. First, as we increase the number of data points, the stationary distribution of the Markov chain will converge weakly to the limiting stationary distribution of Theorem 2.

PROPOSITION 1. *Under Assumptions* 1 *and* 3,

$$\tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}, \mathrm{d}z) \to \tilde{\pi}(\mathrm{d}\tilde{\theta}, \mathrm{d}z)$$

*weakly in* $\mathbb{P}^Y$*-probability as* $T \to \infty$*, where* $\tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}, \mathrm{d}z) = \tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}) \exp(z)\, \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})$ *with* $\tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta})$ *and* $\tilde{g}_T^{\omega}(\mathrm{d}z)$ *as defined in* § 3.2 *and* $\tilde{\pi}(\mathrm{d}\tilde{\theta}, \mathrm{d}z)$ *as defined in* (5).

This result holds because the marginal $\pi_T^{\omega}(\mathrm{d}\theta)$ concentrates around the limiting parameter value $\bar{\theta}$ while the noise converges uniformly to a normal distribution in a neighbourhood around $\bar{\theta}$. The next proposition ensures the stability of the transition and can be proved using similar arguments.

Proposition 2. *Under Assumptions* 1–3*, as* $T \to \infty$ *we have that for any* $f \in \mathrm{BL}(\mathbb{R}^{d+1})$*,*

$$\int \left| \tilde{P}_T^{\omega} f(\theta,\, z) - \tilde{P} f(\theta,\, z) \right| \tilde{\pi}_T^{\omega}(\mathrm{d}\theta, \mathrm{d}z) \to 0$$

*in* $\mathbb{P}^Y$*-probability, where the transition kernels* $\tilde{P}_T^{\omega}$ *and* $\tilde{P}$ *are defined in* (4) *and* (6).

A further requirement to ensure the stability of the transition is that the application of the transition operator should conserve continuity.

Proposition 3. *Under Assumption* 2*, the map* $(\theta,\, z) \mapsto \tilde{P} f(\theta,\, z)$ *is continuous for every* $f \in C_{\mathrm{b}}(\mathbb{R}^{d+1})$.

Theorem 1 now follows from a direct application of Theorem 2, as the assumptions (i), (ii) and (iii) hold by Propositions 1, 2 and 3, respectively.

## 5. Random effects models

### 5.1. *Statistical model and likelihood estimator*

We establish sufficient conditions under which weak convergence of the pseudo-marginal algorithm holds for an important class of latent variable models. Consider the model

$$X_t \sim f(\cdot \mid \theta), \quad Y_t \mid X_t \sim g(\cdot \mid X_t, \theta),$$

where $(X_t)_{t\geqslant 1}$ are independent $\mathbb{R}^k$-valued latent variables, $f(x \mid \theta)$ is a density with respect to Lebesgue measure, and $(Y_t)_{t\geqslant 1}$ are $\mathsf{Y}$-valued observations distributed according to a conditional density $g(y \mid x, \theta)$ with respect to a dominating measure, $\mathsf{Y}$ being a topological space. For observations $Y_{1:T} = y_{1:T}$, the likelihood is

$$p(y_{1:T} \mid \theta) = \prod_{t=1}^{T} p(y_t \mid \theta) = \prod_{t=1}^{T} \int g(y_t \mid x_t, \theta) f(x_t \mid \theta)\, \mathrm{d}x_t.$$

In many scenarios, this likelihood is not available analytically. In order to perform Bayesian inference about the parameter $\theta$, we can use the pseudo-marginal algorithm as it is possible to obtain an unbiased nonnegative estimator of $p(y_{1:T} \mid \theta)$ using importance sampling. Indeed, we can consider $\hat{p}(y_{1:T} \mid \theta, U) = \prod_{t=1}^{T} \hat{p}(y_t \mid \theta, U_t)$ where $U = (U_1, \ldots, U_T)$, $U_t = (U_{t,1}, \ldots, U_{t,N})$, each $U_{t,i}$ is $\mathbb{R}^k$-valued, $N$ denotes the number of Monte Carlo samples, and $\hat{p}(y_t \mid \theta, U_t)$ is important sampling estimator of $p(y_t \mid \theta)$ of the form

$$\hat{p}(y_t \mid \theta, U_t) = \frac{1}{N} \sum_{i=1}^{N} w(y_t, U_{t,i}, \theta), \quad w(y_t, U_{t,i}, \theta) = \frac{g(y_t \mid U_{t,i}, \theta) f(U_{t,i} \mid \theta)}{h(U_{t,i} \mid y_t, \theta)}$$

with $U_{t,i} \sim h(\cdot \mid y_t, \theta)$, where $h(\cdot \mid y_t, \theta)$ is a probability density on $\mathbb{R}^k$ with respect to Lebesgue measure. In this case, the joint density $m_{T,\theta}(u)$ of all the auxiliary variates used to obtain the likelihood estimator is given by the product over $t = 1, \ldots, T$ and $i = 1, \ldots, N$ of $h(u_{t,i} \mid y_t, \theta)$. We will assume in what follows that the true observations are independent and identically distributed samples taken from a probability measure $\mu$, so that the joint data distribution is the product measure $\mathbb{P}^Y(\mathrm{d}\omega) = \prod_{t=1}^{\infty} \mu(\mathrm{d}y_t)$.

### 5.2. *Verifying the assumptions*

The Bernstein–von Mises theorem holds under weak regularity conditions; see van der Vaart (2000, Theorem 10.1) and the Supplementary Material for the case of generalized linear mixed models presented in § 5.3. This ensures that Assumption 1 is satisfied while Assumption 2 is easy to fulfil, selecting for example a multivariate normal proposal with covariance scaling as $1/T$. Assumption 3 is more complicated as it requires the establishment of uniform conditional central limit theorems for $\hat{p}(Y_{1:T} \mid \theta, U)$ in scenarios where $U \sim m_{T,\theta}$ arises from the proposal, so that $Z \sim g_T^{\omega}(\cdot \mid \theta)$, or at stationarity where $U \sim \pi_T^{\omega}(\cdot \mid \theta)$ with

$$\pi_T^{\omega}(u \mid \theta) = \frac{\hat{p}(y_{1:T} \mid \theta, u)}{p(y_{1:T} \mid \theta)} m_{T,\theta}(u),$$

implying that $Z \sim \bar{g}_T^{\omega}(\cdot \mid \theta)$. We let $\sigma^2(y, \theta) = \mathrm{var}\{\bar{w}(y, U_{1,1}, \theta)\}$, $\sigma^2(\theta) = E\{\sigma^2(Y_1, \theta)\}$, with $U_{1,1} \sim h(\cdot \mid y, \theta)$, $Y_1 \sim \mu$ and

$$\bar{w}(Y_t, U_{t,i}, \theta) = \frac{w(Y_t, U_{t,i}, \theta)}{p(Y_t \mid \theta)}. \tag{7}$$

We can show that under the following condition, Assumption 3 holds.

*Assumption* 4. There exist a closed $\varepsilon$-ball $B(\bar{\theta})$ around $\bar{\theta}$ and a function $g$ such that the normalized weight $\bar{w}(y, U_{1,1}, \theta)$ defined in (7) satisfies

$$\sup_{\theta \in B(\bar{\theta})} E\{\bar{w}(y, U_{1,1}, \theta)^{2+\Delta}\} \leqslant g(y)$$

for some $\Delta > 0$, where $U_{1,1} \sim h(\cdot \mid y, \theta)$ and $\mu(g) < \infty$. Additionally, $\theta \mapsto \sigma^2(y, \theta)$ is continuous in $\theta$ on $B(\bar{\theta})$ for all $y \in \mathsf{Y}$.

THEOREM 3. *Under Assumption 4, Assumption 3 is satisfied.*

Theorem 3 strengthens earlier results of Deligiannidis et al. (2018, Theorem 1) which give standard central limit theorems for the error in the loglikelihood estimator.

### 5.3. *Generalized linear mixed models*

A common example of random effects models is the class of generalized linear mixed models (McCulloch & Neuhaus, 2005), where the observation density is a member of the exponential family and the latent variable follows a centred Gaussian distribution. The densities with respect to some dominating measure can be expressed as

$$g(y \mid x, \theta) = \prod_{j=1}^{J} m(y_j) \exp\left[\eta_j(x) T(y_j) - A\{\eta_j(x)\}\right], \quad f(x \mid \theta) = \varphi(x; 0, \tau^2),$$

where $\eta_j(x) = c_j^{\mathrm{T}}\beta + x$, with $c$ being a vector of covariates and $\beta$ the corresponding parameter vector, $A(\eta)$ denotes the log-partition function, and $m(y)$ is a base measure. In the Supplementary Material we show that for many such models the assumptions of Theorem 1 can be verified when using appropriately chosen importance sampling proposals. In particular, we show that Assumption 4 holds and consequently Assumption 3 holds by Theorem 3.

## 6. Efficient implementation of the pseudo-marginal random walk algorithm

### 6.1. *Optimal tuning*

We optimize the performance of the limiting pseudo-marginal chain identified in Theorem 1 as a proxy for the optimization of the original pseudo-marginal chain. We assume that the limiting covariance matrix $\Sigma$ in (3) is the identity matrix $I_d$, with $d$ denoting the parameter dimension. For general covariance matrices, we can use a Cholesky decomposition and a change of variables as in Sherlock et al. (2015) and Nemeth et al. (2016). We denote by $\tilde{P}_{\ell,\sigma}$ the transition kernel (6) using the proposal density

$$q(\theta, \theta') = \varphi(\theta'; \theta, \ell^2 I_d/d).$$

As in Pitt et al. (2012) and Doucet et al. (2015), we propose to minimize $\mathrm{CT}(f, \tilde{P}_{\ell,\sigma})$, as defined in (1), with respect to the noise standard deviation $\sigma$ and, in contrast to those works, also with respect to the scale parameter $\ell$. We restrict attention here to the case where $f(\theta, z) = \theta_1$, the first component of $\theta$, and write $\mathrm{CT}(f, \tilde{P}_{\ell,\sigma}) = \mathrm{CT}(\ell, \sigma)$ in this case. As this criterion is not available in closed form, we simulate the limiting Markov chain initialized in its stationary regime with different noise levels $\sigma$ and scales $\ell$ on a fine grid to obtain empirical estimates of $\mathrm{CT}(\ell, \sigma)$ computed using the overlapping batch mean estimator. This simulation is straightforward as the target and noise distributions in the limiting case are both Gaussian. We then find the approximate minimizer $(\hat{\ell}_{\mathrm{opt}}, \hat{\sigma}_{\mathrm{opt}})$ of $\mathrm{CT}(\ell, \sigma)$ over the grid. This set-up is used for parameter dimension $d$ ranging from 1 to 50. The results are summarized in Table 1.

Table 1 also lists the computing times at these values and the average acceptance probability of the proposal under $\tilde{P}_{\ell,\sigma}$ at stationarity by using five million iterates of the chain. The results are consistent with those in Doucet et al. (2015) and Sherlock et al. (2015). For low dimensions, $1 \leqslant d \leqslant 5$, the ideal Metropolis–Hastings algorithm mixes well and $\hat{\sigma}_{\mathrm{opt}}$ is around 1.1–1.3, as suggested by Doucet et al. (2015), and increases slowly as $d$ increases to the values $(\ell_\infty, \sigma_\infty) = (2.56, 1.81)$ obtained by the diffusion limit (Sherlock et al., 2015). For example, for $d = 50$ we obtain $(\hat{\ell}_{\mathrm{opt}}, \hat{\sigma}_{\mathrm{opt}}) = (2.41, 1.74)$, and the resulting optimal computing time $\mathrm{CT}(\hat{\ell}_{\mathrm{opt}}, \hat{\sigma}_{\mathrm{opt}})$ is close to $\mathrm{CT}(\ell_\infty, \sigma_\infty)$. For lower dimensions, however, the performance in terms of computing time can be improved by reducing $\sigma$ and $\ell$ relative to $\sigma_\infty$ and $\ell_\infty$; see Table 2. We also observe empirically that the cost function $\ell \mapsto \mathrm{CT}(\ell, \sigma)$ is fairly flat, as noticed by Sherlock et al. (2015) in the limiting case.

### 6.2. *Implementation*

We now show how to exploit the results of the previous subsection in practice to design an efficient implementation of the pseudo-marginal algorithm. Using a preliminary run, we compute estimates $\hat{\theta}$ and $\hat{\Sigma}$ of the posterior mean and posterior covariance matrix. For the parameter dimension $d$, we choose $\ell$ according to Table 1 and use a Gaussian random walk proposal with covariance matrix $\hat{\ell}_{\mathrm{opt}}^2 \hat{\Sigma}/d$. Finally, we select the number $N$ of Monte Carlo samples such that the sample standard deviation of the loglikelihood estimate at $\hat{\theta}$ matches the optimal value $\hat{\sigma}_{\mathrm{opt}}$

Table 1. *Optimal values for the scaling $\ell$ and the noise $\sigma$, along with the associated computing time and average acceptance probability; reported values are the mean and standard deviation (in parentheses) of the minimizers over* 10 *runs*

| Dimension $d$ | $\hat{\ell}_{\text{opt}}$ | $\hat{\sigma}_{\text{opt}}$ | $\text{CT}(\hat{\ell}_{\text{opt}}, \hat{\sigma}_{\text{opt}})$ | $\text{pr}_{\text{acc}}(\hat{\ell}_{\text{opt}}, \hat{\sigma}_{\text{opt}})$ |
|---|---|---|---|---|
| $d = 1$ | 2.05 (0.25) | 1.16 (0.07) | 8.47 | 25.73% |
| $d = 2$ | 1.97 (0.14) | 1.21 (0.06) | 12.71 | 22.92% |
| $d = 3$ | 2.11 (0.07) | 1.24 (0.05) | 16.79 | 19.97% |
| $d = 5$ | 2.17 (0.12) | 1.30 (0.05) | 23.18 | 17.35% |
| $d = 10$ | 2.20 (0.08) | 1.44 (0.05) | 37.93 | 14.27% |
| $d = 15$ | 2.33 (0.08) | 1.50 (0.00) | 53.43 | 12.07% |
| $d = 20$ | 2.34 (0.10) | 1.54 (0.05) | 65.62 | 11.44% |
| $d = 30$ | 2.36 (0.11) | 1.61 (0.03) | 90.46 | 10.41% |
| $d = 50$ | 2.41 (0.10) | 1.74 (0.05) | 136.38 | 8.66% |

Table 2. *Comparison of the computing times for different noise levels; $\hat{\sigma}_{\text{opt}}$ denotes the minimizer of the estimated integrated autocorrelation time, as shown in Table* 1

| Dimension $d$ | $\text{CT}(\ell_\infty, \hat{\sigma}_{\text{opt}})$ | $\text{CT}(\ell_\infty, \sigma = 1.2)$ | $\text{CT}(\ell_\infty, \sigma_\infty)$ |
|---|---|---|---|
| $d = 1$ | 9.04 (0.25) | 9.05 (0.21) | 17.10 (1.34) |
| $d = 2$ | 13.48 (0.32) | 13.37 (0.28) | 22.45 (0.81) |
| $d = 3$ | 17.63 (0.28) | 17.43 (0.26) | 26.71 (0.64) |
| $d = 5$ | 24.38 (0.44) | 24.72 (0.31) | 34.14 (0.88) |
| $d = 10$ | 40.17 (0.71) | 41.60 (0.24) | 47.08 (1.03) |
| $d = 15$ | 53.69 (0.72) | 58.01 (0.50) | 59.08 (0.79) |
| $d = 20$ | 67.15 (0.53) | 74.34 (0.36) | 71.41 (1.48) |
| $d = 30$ | 91.36 (0.95) | 106.08 (0.34) | 93.73 (1.08) |
| $d = 50$ | 136.49 (1.18) | 167.83 (0.75) | 135.92 (1.27) |

listed in Table 1. This approach is similar to the one taken in Sherlock et al. (2015), except for the dimension dependence of the recommended parameters $(\hat{\ell}_{\text{opt}}, \hat{\sigma}_{\text{opt}})$.

## 7. SIMULATION STUDY: RANDOM EFFECTS MODEL

In this section we illustrate how the guidelines derived from the limiting pseudo-marginal chain compare to a practical implementation of the pseudo-marginal algorithm. We consider a logistic mixed effects model applied to a real dataset. Mixed models are popular in econometrics, survey analysis and medical statistics, among other fields, and are often used to describe heterogeneity between groups. Here we consider a subset of a cohort study of Indonesian preschool children. This dataset was previously analysed by Zeger & Karim (1991) using Bayesian mixed models. It contains 1200 observations of 275 children. We model the probability of a respiratory infection based on the following covariates: age, sex, height, an indicator for vitamin deficiency, an indicator for subnormal height and two seasonal components. Including the intercept, there are eight covariates. Cluster effects due to repeated measurements from the same children are modelled with individual random intercepts. In this case the linear predictor of a regression model based on covariates $c_{t,j}$ $(t = 1, \ldots, T; j = 1, \ldots J)$ reads $\eta_{t,j} = c_{t,j}^{\mathrm{T}}\beta + X_t$, where $X_t \sim \mathcal{N}(0, \tau)$ denotes the random intercept for child $t = 1, \ldots, T$ and $\beta$ the regression parameters. For every child we have an observation vector $y_t = (y_{t,1}, \ldots, y_{t,J}) \in \{0, 1\}^J$. The unknown parameter is $\theta = (\beta, \tau) \in \mathbb{R}^d$, where $d = 9$. The observations are assumed to be conditionally independent

Table 3. *Results of the simulation study for $N$ particles: standard deviation $\hat{\sigma}$ of the loglikelihood estimator at the mean, average integrated autocorrelation time* $\widehat{\text{IAT}}$*, and average acceptance probability* $\widehat{\text{pr}}_{\text{acc}}$ *for the pseudo-marginal kernel and the limiting kernel* $\tilde{P}_{\ell,\hat{\sigma}}$ *for* $\ell = 2.2$

| $N$ | $\hat{\sigma}$ | $\widehat{\text{IAT}}$ | $\widehat{\text{pr}}_{\text{acc}}$ | $\widehat{\text{IAT}}(\tilde{P}_{\ell=2.2,\sigma=\hat{\sigma}})$ | $\widehat{\text{pr}}_{\text{acc}}(\tilde{P}_{\ell=2.2,\sigma=\hat{\sigma}})$ |
|---|---|---|---|---|---|
| 12 | 2.00 | 140.22 | 8.93% | 162.57 | 7.67% |
| 15 | 1.76 | 112.06 | 10.70% | 121.70 | 9.93% |
| 18 | 1.63 | 98.69 | 12.30% | 94.14 | 11.73% |
| 21 | 1.46 | 72.42 | 13.93% | 72.31 | 14.00% |
| 24 | 1.34 | 66.29 | 15.10% | 64.45 | 15.55% |
| 27 | 1.29 | 61.95 | 16.08% | 58.08 | 16.39% |
| 30 | 1.22 | 58.70 | 16.85% | 54.12 | 17.52% |
| 33 | 1.16 | 52.39 | 17.77% | 50.26 | 18.16% |

given the random effects and are modelled by

$$g(y_t \mid x_t, \theta) = \prod_{j=1}^{J} \frac{\exp(y_{t,j}\eta_{t,j})}{1+\exp(\eta_{t,j})}, \quad f(x_t \mid \theta) = \varphi(x_t; 0, \tau) \quad (t = 1, \ldots, T).$$

Inference in mixed effects models often aims to find the population effects, so one is interested in integrating out the random effects. Since the marginal likelihood contains intractable integrals, this model lends itself to the pseudo-marginal approach. We obtain an unbiased estimator of the marginal likelihood by estimating the integrals using an importance sampling estimator

$$h(u \mid y_t, \theta) = \varphi(u; \hat{x}_t, \tau_q^2), \quad \hat{x}_t = \arg\max_{x_t} g(y_t \mid x_t, \theta) f(x_t \mid \theta)$$

with proposal variance $\tau_q > 0$. More details on importance sampling for mixed effects models are provided in the Supplementary Material, where we also show that Assumption 4 is satisfied in the present example. For the covariate parameters we assume a diffuse Gaussian prior, and the variance of the random effects is assigned an inverse gamma prior. We run a pseudo-marginal algorithm with a Gaussian random walk proposal for 500 000 iterations. The covariance of the proposal is set equal to the posterior covariance of the parameters estimated in a preliminary run and scaled by $\ell^2/d = (2.2)^2/9$. We compare the average integrated autocorrelation time and the acceptance rate with that of the limiting chain using the same $\ell = 2.2$ and $\sigma = \hat{\sigma}$, the average being defined as $\widehat{\text{IAT}}(P_T^{\omega}) = \sum_{i=1}^{d} \text{IAT}(f_i, P_T^{\omega})$ where $f_i(\theta, z) = \theta_i$ is the $i$th parameter component. Here, $\hat{\sigma}$ is the standard deviation of the loglikelihood estimator obtained using 10 000 samples of the marginal likelihood evaluated at an estimate $\hat{\theta}$ of the posterior mean. The results are summarized in Table 3. For a given number of particles $N$, we report the associated estimate of the noise in the loglikelihood estimator, the average integrated autocorrelation time, and the average acceptance rate.

The average integrated autocorrelation time and the acceptance rate are very close to those of the limiting algorithm. This is visualized in Fig. 1, where these quantities are plotted against the number of particles $N$. The computing time of the pseudo-marginal algorithm targeting the posterior, $\widehat{\text{CT}}(P_T^{\omega}) = \widehat{\text{IAT}}(P_T^{\omega})/\hat{\sigma}^2$, and the computing time of the limiting algorithm, $\widehat{\text{CT}}(\tilde{P}_{\ell,\sigma})$, are both optimized for $\hat{\sigma} = 1.46$, as expected from Table 1. In this example, the limiting kernel captures very well the behaviour of the pseudo-marginal algorithm for large datasets, and Table 1 thus provides useful guidelines on how to tune this scheme.

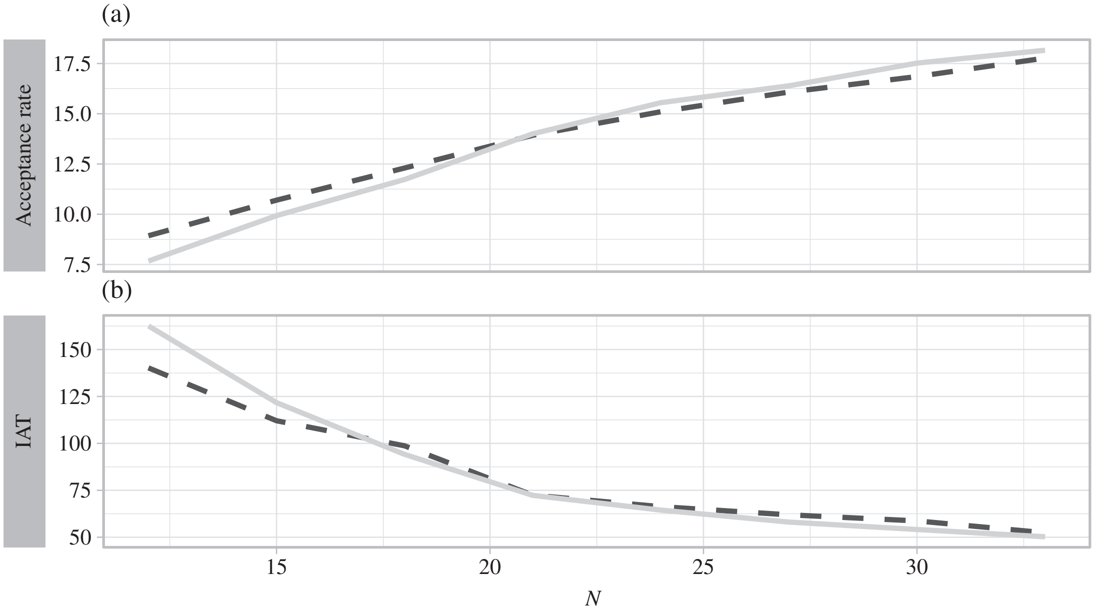


Fig. 1. (a) Average acceptance rate and (b) average integrated autocorrelation time for the pseudo-marginal algorithm (black dashed line) and the limiting transition kernel $\tilde{P}_{\ell,\hat{\sigma}}$ for $\ell = 2.2$ (grey solid line) as functions of $N$, the number of particles.

## Acknowledgement

We are thankful to the reviewers for their comments, which have helped us to improve the paper significantly. Schmon and Doucet were supported by the U.K. Engineering and Physical Sciences Research Council.

## Supplementary material

Supplementary material available at *Biometrika* online includes proofs of all the propositions and theorems, as well as a set of generalized linear mixed models for which all the assumptions hold. It also contains a short review of weak convergence of random measures and some further simulation studies, including a three-dimensional Lotka–Volterra model.

## References

Andrieu, C., Doucet, A. & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods (with Discussion). *J. R. Statist. Soc.* B **72**, 269–342.

Andrieu, C. & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37**, 697–725.

Andrieu, C. & Vihola, M. (2015). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Prob.* **25**, 1030–77.

Andrieu, C. & Vihola, M. (2016). Establishing some order amongst exact approximations of MCMCs. *Ann. Appl. Prob.* **26**, 2661–96.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164**, 1139–60.

Bérard, J., Del Moral, P. & Doucet, A. (2014). A lognormal central limit theorem for particle approximations of normalizing constants. *Electron. J. Prob.* **19**, 1–28.

Crauel, H. (2003). *Random Probability Measures on Polish Spaces*. Boca Raton, Florida: CRC Press.

Deligiannidis, G., Doucet, A. & Pitt, M. K. (2018). The correlated pseudomarginal method. *J. R. Statist. Soc.* B **80**, 839–70.

Doucet, A., Pitt, M. K., Deligiannidis, G. & Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika* **102**, 295–313.

Kallenberg, O. (2006). *Foundations of Modern Probability*. New York: Springer.

Kleijn, B. J. K. & Van der Vaart, A. W. (2012). The Bernstein-von-Mises theorem under misspecification. *Electron. J. Statist.* **6**, 354–81.

Lin, L., Liu, K. & Sloan, J. (2000). A noisy Monte Carlo algorithm. *Phys. Rev.* D **61**, 074505.

McCulloch, C. E. & Neuhaus, J. M. (2005). Generalized linear mixed models. In *Encyclopedia of Biostatistics*, vol. 4. Chichester: John Wiley & Sons.

Nemeth, C., Sherlock, C. & Fearnhead, P. (2016). Particle Metropolis-adjusted Langevin algorithms. *Biometrika* **103**, 701–17.

Pitt, M. K., dos Santos Silva, R., Giordani, P. & Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *J. Economet.* **171**, 134–51.

Roberts, G., Gelman, A. & Gilks, W. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Prob.* **7**, 110–20.

Roberts, G. O. & Rosenthal, J. S. (2014). Minimising MCMC variance via diffusion limits, with an application to simulated tempering. *Ann. Appl. Prob.* **24**, 131–49.

Sherlock, C., Thiery, A. H. & Lee, A. (2017). Pseudo-marginal Metropolis–Hastings sampling using averages of unbiased estimators. *Biometrika* **104**, 727–34.

Sherlock, C., Thiery, A. H., Roberts, G. O. & Rosenthal, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.* **43**, 238–75.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Zeger, S. L. & Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. Am. Statist. Assoc.* **86**, 79–86.

[*Received on* 10 *July* 2018. *Editorial decision on* 29 *January* 2020]

*Biometrika* (2017), **103**, 1, *pp*. 1–48
*Printed in Great Britain* Advance Access publication on 31 July 2016

# Supplementary Material to 'Large Sample Asymptotics of the Pseudo-Marginal Method'

BY S. M. SCHMON, G. DELIGIANNIDIS, A. DOUCET
*Department of Statistics, University of Oxford*
*24-29 St Giles', Oxford OX1 3LB*
schmon@stats.ox.ac.uk deligiannidis@stats.ox.ac.uk doucet@stats.ox.ac.uk

AND M. K. PITT
*Department of Mathematics, King's College London*
*Strand, London WC2R 2LS*
michael.pitt@kcl.ac.uk

SUMMARY

This supplementary material contains the proofs to all theorems and propositions, some background material and additional simulation studies. Section S1 includes a brief survey of weak convergence results for random probability measures on Polish spaces which play an important role in this article. We have not been able to find some of the precise statements we require in the literature so we present their proofs here without any claim of originality. Sections S2 and S3 provide the proofs for sections 4 and 5, respectively. Finally, section S4 includes some additional numerical examples: a toy example and a Lotka-Volterra model where the likelihood is estimated using a particle filter as opposed to importance sampling.

## S1. RANDOM MEASURES AND WEAK CONVERGENCE ON POLISH SPACES

### S1·1. *Weak Convergence*

Let $S$ be a Polish space, endowed with the Borel $\sigma$-algebra $\mathcal{B}(S)$. We denote $d$ the metric inducing the topology on $S$ and $\mathcal{P}(S)$ the space of Borel probability measures on $S$. In the following, we will only consider (random) probability measures in $\mathcal{P}(S)$ unless stated otherwise.

DEFINITION 1 (*Weak convergence*). *A sequence of probability measures $(\mu_n)_{n\geqslant 1}$ converges weakly to a probability measure $\mu$, denoted $\mu_n \rightsquigarrow \mu$, if for all $f \in C_b(S)$*

$$\mu_n(f) \to \mu(f) \quad \text{as } n \to \infty, \tag{1.1}$$

*where $C_b(S)$ is the set of bounded continuous real-valued functions of domain $S$.*

The set of test functions generating this topology can be restricted to bounded continuous functions $f : S \to [0, 1]$ or bounded Lipschitz functions, see for example Crauel (2003, Lemma A.1 and Theorem A.2). The topology of weak convergence can be metrized using the bounded Lipschitz metric which is given for $\mu, \nu \in \mathcal{P}(S)$ by

$$d_{\mathrm{BL}}(\mu, \nu) = \sup \{|\mu(f) - \nu(f)| \,;\, f \in \mathrm{BL}(S), \|f\|_{\mathrm{BL}} \leq 1\}, \tag{1.2}$$

see for example Dudley (2002, Proposition 11.3.2). Here, the set $\mathrm{BL}(S)$ denotes the set of bounded Lipschitz functions and we follow Pollard (2002) by defining the norm

$$\|f\|_{\mathrm{BL}} = \max \{\|f\|_{\mathrm{L}}, 2\|f\|_\infty\}, \tag{1.3}$$

where

$$\|f\|_{\mathrm{L}} = \sup_{x,y:x\neq y} \frac{|f(x)-f(y)|}{d(x,y)} \quad \text{and} \quad \|f\|_\infty = \sup_x |f(x)|. \tag{1.4}$$

This definition gives us the inequality

$$|f(x)-f(y)| \leq \|f\|_{\mathrm{BL}} \left[\min\{1, d(x,y)\}\right] \tag{1.5}$$

for every $x, y$.

## S1·2. *Weak Convergence of Random Measures*

We recall here some facts about random probability measures. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space. We equip the product space $\Omega \times S$ with the product $\sigma$-algebra, $\mathcal{F} \otimes \mathcal{B}(S)$.

DEFINITION 2 (*Random probability measure*). *A* random probability measure *is a map* $\mu\colon \Omega \times \mathcal{B}(S) \to [0,1]$ *such that for every* $B \in \mathcal{B}(S)$ *the map* $\omega \mapsto \mu(\omega, B) = \mu^\omega(B)$ *is measurable while* $\mu(\omega, \cdot) \in \mathcal{P}(S)$ *for almost every* $\omega \in \Omega$.

For all bounded and measurable functions $g\colon \Omega \times S \to \mathbb{R}$, the assignment $\omega \mapsto \int_S g(\omega, x)\mu^\omega(\mathrm{d}x)$ is measurable (see, for example, Crauel, 2003, Proposition 3.3) and thus, for random measures, the map $\omega \mapsto \mu^\omega(f)$ is a random variable. As a consequence we have that $\mu^\omega\colon \Omega \to \mathcal{P}(S)$ is a Borel measurable map. Conversely, it can be shown that any random element of $[\mathcal{P}(S), \mathcal{B}\{\mathcal{P}(S)\}]$ fulfils the condition set out in Definition 1, see (Crauel, 2003, Remark 3.20 (i)) or (Kallenberg, 2006, Lemma 1.37) for details.

DEFINITION 3 (*Weak convergence of random measures*). *A sequence of random probability measures* $(\mu_n^\omega)_{n\geqslant 1}$ *converges weakly almost surely to a probability measure* $\mu$, *denoted* $\mu_n^\omega \rightsquigarrow_{a.s.} \mu$, *if*

$$\mathbb{P}\left(\omega \in \Omega : \ \mu_n^\omega \rightsquigarrow \mu\right) = 1. \tag{1.6}$$

*Further, we say that* $(\mu_n^\omega)_{n\geqslant 1}$ *converges weakly in probability, denoted* $\mu_n^\omega \rightsquigarrow_{\mathbb{P}} \mu$, *if every subsequence contains a further subsequence which converges weakly almost surely.*

One can easily verify that the above definition of almost sure weak convergence, respectively weak convergence in probability, is equivalent to $\rho(\mu_n^\omega, \mu) \to 0$ almost surely, respectively in probability, for some metric $\rho$ on $\mathcal{P}(S)$ metrizing weak convergence, e.g., the bounded Lipschitz metric (1.2), see for example Theorem 1.

*Remark* 1 (Measurability of probability metric). *As already mentioned above, for any random measure the map* $\omega \mapsto \mu^\omega$ *is measurable with respect to the Borel* $\sigma$*-algebra* $\mathcal{B}\{\mathcal{P}(S)\}$. *Moreover, any metric* $\rho$ *inducing the weak topology on* $\mathcal{P}(S)$ *is trivially continuous in its first argument and hence the map* $\mu^\omega \mapsto \rho(\mu^\omega, \nu)$ *for some fixed measure* $\nu$ *is measurable with respect to the Borel* $\sigma$*-algebra* $\mathcal{B}(\mathbb{R})$. *This implies (Borel) measurability of the map* $\omega \mapsto \rho(\mu^\omega, \nu)$ *for a non-random measure* $\nu$.

In light of the definition of weak convergence (1.1) it is natural to ask whether almost sure weak convergence holds if

$$\mu_n^\omega(f) \xrightarrow{\text{a.s.}} \mu(f) \qquad \text{for all} \qquad f \in C_b(S), \tag{1.7}$$

and similarly whether weak convergence in probability holds if

$$\mu_n^\omega(f) \xrightarrow{\mathbb{P}} \mu(f) \qquad \text{for all} \qquad f \in C_b(S). \tag{1.8}$$

In many practical applications, it appears easier to check (1.7) rather than (1.6), similarly checking (1.8) appears easier than having to check that every subsequence of $(\mu_n^\omega)_{n\geqslant 1}$ contains a subsequence which converges weakly almost surely. Relating those statements is inconvenienced by the fact that weak convergence is usually checked using an uncountable convergence determining class of functions, e.g., the space of bounded continuous functions. However, we show here that these equivalences hold true for Polish spaces; see Theorem 1 below.

Almost sure weak convergence can be shown using the existence of a countable convergence determining subclass $\mathcal{C} \subset \mathrm{BL}(S) \subset C_b(S)$. Considering subsequences and using a diagonal argument we can show the equivalence of the statement also holds if almost sure convergence is replaced by convergence in probability. For the purposes of this paper we confine our attention to weak convergence in probability. To prove the statements above we first need an auxiliary result, which also appeared in Sweeting (1989, Lemma 4).

PROPOSITION 1. *Suppose $A$ is a countable set and consider random variables $X_n(a)\colon \Omega \to \mathbb{R}$ indexed by $a \in A$ and $n \in \mathbb{N}$. Moreover, assume that for every $a \in A$ the sequence $\{X_n(a)\}_{n\geqslant 1}$ converges to $X(a)$ in probability, i.e.,*

$$X_n(a) \xrightarrow{\mathbb{P}} X(a) \quad \forall a \in A.$$

*Then there exists a subsequence $N' \subset \mathbb{N}$ such that along $N'$*

$$\mathbb{P}\{\omega : X_n(a) \to X(a) \quad \forall a \in A\} = 1.$$

*Proof.* Choose $a_1 \in A$. Since we have $X_n(a_1) \xrightarrow{\mathbb{P}} X(a_1)$ we can extract a subsequence $n_{1,1}, n_{1,2}, \ldots$ such that

$$\left\{X_{n_{1,1}}(a_1), X_{n_{1,2}}(a_1), X_{n_{1,3}}(a_1), \ldots\right\}$$

converges almost surely. Pick now $a_2 \in A$, we can now extract a further subsequence

$$\left\{X_{n_{2,1}}(a_2), X_{n_{2,2}}(a_2), X_{n_{2,3}}(a_2), \ldots\right\}$$

along which we have almost sure convergence. We can iterate this procedure to get another subsequence

$$\left\{X_{n_{3,1}}(a_3), X_{n_{3,2}}(a_3), X_{n_{3,3}}(a_3), \ldots\right\}.$$

Along the subsequence $N' = \left(n_{1,1}, n_{2,2}, n_{3,3}, ...\right)$, we have almost sure convergence of $X_{n'}(a) \to X(a)$ for all $a \in A$. □

The existence of a countable convergence determining class for Polish spaces is guaranteed by the following Proposition. The proof is adapted from Berti et al. (2006, Theorem 2.2).

PROPOSITION 2. *Consider $\mathcal{P}(S)$ equipped with the Borel $\sigma$-algebra generated by the topology of weak convergence. There exists a countable convergence determining subclass $\mathcal{C} \subset \mathrm{BL}(S)$.*

*Proof.* Take a countable set $\{s_1, s_2, \ldots\}$ dense in $S$ and let $H = [0, 1]^{\mathbb{N}}$ be the Hilbert cube. For $x \in S$, define the map $h\colon S \to H$ by

$$h(x) = \{d(x, s_1) \wedge 1, d(x, s_2) \wedge 1, \ldots\}.$$

We can equip $H$ with the topology of coordinate wise convergence. Writing $u = (u_1, u_2, \ldots)$ and $v = (v_1, v_2, \ldots)$ for elements $u, v \in H$, this topology is induced by the metric

$$\alpha(u, v) = \sum_{i=1}^{\infty} \frac{|u_i - v_i|}{2^i}.$$

The Hilbert cube $H$ is compact by Tychonoff's Theorem (see for example Dudley, 2002, Theorem 2.2.8.), $h$ is a homeomorphism from $S$ to $h(S)$ (Borkar, 1991, Theorem A.1.1.) and its closure $\overline{h(S)} \subset H$ is compact. For $\mu \in \mathcal{P}(S)$ denote $\nu = \mu \circ h^{-1}$ the image measure on $h(S)$.

Note that any Lipschitz continuous function on $h(S)$ can be extended to $\overline{h(S)}$ without increasing its norm (Dudley, 2002, Proposition 11.2.3.). By the Arzelà–Ascoli theorem, the sets $B_n = [f \in \mathrm{BL}\{\overline{h(S)}\} : \|f\|_{\mathrm{BL}} \leq n]$ are compact and thus separable under the $\|\cdot\|_\infty$-norm. Therefore $\mathrm{BL}\{\overline{h(S)}\} = \bigcup_{n=1}^{\infty} B_n$ is separable under the $\|\cdot\|_\infty$-norm and so is $\mathrm{BL}\{h(S)\}$. Hence, we can pick a countable set $\mathcal{D}$ which is dense in $\mathrm{BL}\{h(S)\}$. Defining $\mathcal{C} = \{g \circ h : g \in \mathcal{D}\}$ we have $\mathcal{C} \subset \mathrm{BL}(S)$ since for all $x, y \in S$ and $i \in \mathbb{N}$

$$|d(x, s_i) \wedge 1 - d(y, s_i) \wedge 1| \leq d(x, y)$$

and thus

$$|g \circ h(x) - g \circ h(y)| \leq L_g \alpha\{h(x), h(y)\} = L_g \sum_{i=1}^{\infty} \frac{|d(x, s_i) \wedge 1 - d(y, s_i) \wedge 1|}{2^i} \leq L_g d(x, y),$$

where $L_g$ denotes the Lipschitz constant of the function $g$.

Now assume that $\mu_n(f) \to \mu(f)$ for all $f \in \mathcal{C}$. Then by a change of variable

$$\int_S f \, d\mu_n = \int_S g \circ h \, d\mu_n = \int_{h(S)} g \, d\nu_n \to \int_{h(S)} g \, d\nu$$

for all $g \in \mathcal{D}$. Since $\mathcal{D}$ is dense in $\mathrm{BL}\{h(S)\}$ with respect to the $\|\cdot\|_\infty$-norm we have convergence for all bounded Lipschitz functions and thus $\nu_n \rightsquigarrow \nu$. By continuity of $h^{-1}$ we also have convergence $\mu_n \rightsquigarrow \mu$.□

Equipped with these results we can now prove some equivalences which facilitate the verification of weak convergence of random probability measures in the sense introduced above. We will prove the following statements only for convergence in probability. The modifications for almost sure convergence are obvious.

THEOREM 1. *Let $(\mu_n^\omega)_{n\geqslant 1}$ be a sequence of random probability measures and $\mu$ a probability measure. Then the following statements are equivalent*

*(i)* $d_{\mathrm{BL}}(\mu_n^\omega, \mu) \xrightarrow{\mathbb{P}} 0,$
*(ii)* $\mu_n^\omega \rightsquigarrow_{\mathbb{P}} \mu$
*(iii)* $\mu_n^\omega(f) \xrightarrow{\mathbb{P}} \mu(f)$ *for all* $f \in C_b(S)$
*(iv)* $\mu_n^\omega(f) \xrightarrow{\mathbb{P}} \mu(f)$ *for all* $f \in \mathrm{BL}(S)$.

*The same results hold if convergence in probability is replaced by almost sure convergence throughout.*

*Proof.* The equivalence $(i) \Leftrightarrow (ii)$ is immediate since $d_{\mathrm{BL}}$ metrizes weak convergence. The implications $(ii) \Rightarrow (iii) \Rightarrow (iv)$ are trivial. To show $(iv) \Rightarrow (ii)$, note that by Proposition 2 there exists a countable convergence determining subclass $\mathcal{C} \subset \mathrm{BL}(S)$. By virtue of Proposition 1 there exists a subsequence $(n_1, n_2, \ldots)$ such that for all $g \in \mathcal{C}$

$$\mu_{n_k}^\omega(g) \xrightarrow{\text{a.s.}} \mu(g) \qquad \text{as } k \to \infty.$$

Now, given $(n_k)_{k\in\mathbb{N}}$ define

$$A(g) = \left\{\omega \in \Omega : \mu_{n_k}^\omega(g) \longrightarrow \mu(g) \quad \text{as} \quad k \to \infty\right\}.$$

We have $\mathbb{P}\{A(g)\} = 1$ for all $g \in \mathcal{C}$ and for $\bigcap_{g\in\mathcal{C}} A(g) = A \in \mathcal{B}(S)$ we find $\mathbb{P}(A) = 1$. Since we can apply this reasoning to any subsequence we always find a further subsequence such that $(\mu_{n_{k_j}}^\omega)$ converges almost surely. See also Sweeting (1989, Theorem 9) and Berti et al. (2006, Theorem 2.2). □

*Remark* 2. If the random measure is induced by a regular conditional distribution, i.e., let $(\mu_n^\omega)_{n\geqslant 1}$ denote a sequence of transition kernels such that

$$\mu_n^\omega(\cdot) = \mathbb{P}(X_n \in \cdot \mid \mathcal{F}_n)(\omega) \qquad \mathbb{P} - a.s.$$

for some filtration $(\mathcal{F}_n)_{n\geqslant 1}$, we have

$$\int f(x)\mu_n^\omega(\mathrm{d}x) = E\left\{f(X_n) \mid \mathcal{F}_n\right\}(\omega) \qquad \mathbb{P} - a.s.$$

and thus equivalently to $\mu_n \rightsquigarrow_{\mathbb{P}} \mu$ then we can write

$$E\{f(X_n) \mid \mathcal{F}_n\} \xrightarrow{\mathbb{P}} E\{f(X)\}, \tag{1.9}$$

where $X \sim \mu$. For brevity we will also use the notation $X_n \mid \mathcal{F}_n \rightsquigarrow_{\mathbb{P}} \mu$ instead of (1.9).

### S1·3. *Product Spaces*

We address here the setting where the spaces are of the form $S^k = S \times S \times \cdots \times S$ or $S^{\mathbb{N}} = S \times S \times \ldots$. We will equip these product spaces with the product topology and the respective Borel $\sigma$-algebra. The following lemma is helpful to characterize weak convergence in probability in this context.

LEMMA 1. *For fixed $k$, let $(\mu_n^\omega)_{n\geqslant 1}$ denote random measures on $S^k$ and $\mu$ a non-random measure on $S^k$. Then the following are equivalent*

*(i)*

$$\mu_n^\omega \rightsquigarrow_{\mathbb{P}} \mu,$$

*(ii)*

$$\mu_n^\omega(f) \xrightarrow{\mathbb{P}} \mu(f)$$

for all $f \in C_b(S^k)$.

*(iii)*

$$\int_{S^k} \prod_{i=1}^k f_i(x_i)\mu_n^\omega(\mathrm{d}x_1 \ldots \mathrm{d}x_k) \xrightarrow{\mathbb{P}} \int_{S^k} \prod_{i=1}^k f_i(x_i)\mu(\mathrm{d}x_1 \ldots \mathrm{d}x_k)$$

for all $f_1, \ldots f_k \in C_b(S)$.

*(iv)*

$$\int_{S^k} \prod_{i=1}^k f_i(x_i)\mu_n^\omega(\mathrm{d}x_1 \ldots \mathrm{d}x_k) \xrightarrow{\mathbb{P}} \int_{S^k} \prod_{i=1}^k f_i(x_i)\mu(\mathrm{d}x_1 \ldots \mathrm{d}x_k)$$

for all $f_1, \ldots f_p \in \mathrm{BL}(S)$.

*Proof.* The implications $(i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (iv)$ are trivial. Thus, we only need to show $(iv) \Rightarrow (i)$. We now by Proposition 2 that there exists a countable convergence determining class $\mathcal{C} \subset \mathrm{BL}(S)$, so we can assume $f_1, f_2, \ldots \in \mathcal{C}$. Without loss of generality we can assume $\|f_i\|_\infty \leq 1$ for all $i$ and $1 \in \mathcal{C}$. Then we have that for every $i \in \{1, \ldots, k\}$ the marginal of the $i$th coordinate, denoted $\mu_{n,i}^\omega$, converges to $\mu_i$ weakly in probability, i.e. for all $i$ and all $f_i \in \mathcal{C}$ we have

$$\int_S f_i(x)\mu_{n,i}^\omega(\mathrm{d}x_i) \xrightarrow{\mathbb{P}} \int_S f_i(x)\mu_i(\mathrm{d}x_i).$$

Now by Proposition 1 for every $i \in \{1, \ldots, k\}$ every subsequence $N \subset \mathbb{N}$ contains a further subsequence $N' \subset N$ such that we have convergence almost sure convergence for all $g \in \mathcal{C}$, i.e. denoting

$$A_i := \left\{\omega \in \Omega : \int_S g(x_i)\mu_{n',i}^\omega(\mathrm{d}x_i) \longrightarrow \int_S g(x_i)\mu_i(\mathrm{d}x_i) \quad \text{for all } g \in \mathcal{C}\right\}$$

we have $P(A_i) = 1$. We can extract a further subsequence $N'' \subset N'$ such that along $N''$ we have convergence almost surely for all $i$ and all $g$ and thus for $\omega \in A := \cap_{i=1}^k A_i$ the sequence $\left\{\mu_n^\omega; n \in N''\right\}$ is tight, since $\left\{\mu_{n,i}^\omega; n \in N''\right\}$ is tight for every $i$ (see Ethier & Kurtz, 2005, Chapter 3 Proposition 2.4.). We can conclude that for every such $\omega$ every subsequence of $(\mu_n^\omega)_{n\geqslant 1}$ has a further subsequence that converges. It remains to show that the functions of the form $\prod_{i=1}^k f_i$ are measure determining. However, by Ethier & Kurtz (2005, Chapter 2 Proposition 4.6.) if $\mathcal{C}$ is measure determining on $S$ then so is the product for $S^k$.□

If $S = \mathbb{R}^k$ for some $k \in \mathbb{N}$ we can check weak convergence in probability by considering moment generating functions. The following result is shown by Sweeting (1989, Corollary 3); see also Castillo & Rousseau (2015, Lemma 1).

PROPOSITION 3. *Let $\left(\mu_n^\omega\right)_{n\geqslant 1}$ be a sequence of random probability measures and assume there exists $u_0 > 0$ such that for all $n \in \mathbb{N}$ the moment generating functions*

$$m_n(u, \omega) = \int \exp\left(u^\mathsf{T} x\right) \mu_n^\omega(\mathrm{d}x)$$

*exist for $|u| < u_0$ then $\mu_n^\omega \rightsquigarrow_{\mathbb{P}} \mu$ if and only if for every $u \in \mathbb{R}^k$*

$$m_n(u, \cdot) \xrightarrow{\mathbb{P}} m(u, \cdot) = \int \exp\left(u^\mathsf{T} x\right) \mu^\omega(\mathrm{d}x).$$

*Proof.* This can be seen by considering the class of functions of the form $f_u(x) = \exp(u^\mathrm{T} x)$ for $u \in \mathbb{Q}, \quad |u| < u_0$ and showing that they form a countable convergence determining class, see Sweeting (1989, Corollary 3). Consider the case $k = 1$ and a sequence of measures $(\mu_n)_{n\geqslant 1}$ and $\mu$ such that

$$m_n(u) = \int e^{ux} \mu_n(\mathrm{d}x) \to m(u) = \int e^{ux} \mu(\mathrm{d}x).$$

Denote a compact set $K = [-c, c]$. Then by the Markov inequality

$$\mu_n\left(K^\complement\right) = \int_{|x|\geq c} \mu_n(dx) \leq \frac{m_n(u_0)}{e^{u_0 c}}$$

and $m_n(u_0) \to m(u_0)$. Hence, $\mu_n(K^\complement)$ is bounded and we can find $c$ such that $\sup_n \mu_n(K^\complement) < \epsilon$ and $(\mu_n)_{n\geqslant 1}$ is tight. By continuity the $f_u$ are measure determining so we can conclude that the limit is unique. For $k > 1$ we can use the same argument to show that the marginals are tight, see the proof of Lemma 1. □

Lemma 1 can be readily extended to countably infinite product spaces by considering convergence of the finite dimensional distribution. Let us therefore denote $\mu \circ \pi_k^{-1} \colon S^\mathbb{N} \to S^k; k \in \mathbb{N}$ the canonical projections. For non-random measures, it is well-known that convergence of the projections already implies convergence on the whole of $S^\mathbb{N}$ (Billingsley, 1999, Example 2.6). Since there are countably many such projections, we can apply the reasoning of Proposition 1 to conclude that for checking $\mu_n^\omega \rightsquigarrow_{\mathbb{P}} \mu$ on $S^\mathbb{N}$ we just need to show

$$\int_{S^k} \prod_{i=1}^k f_i(x_i) \mu_n^\omega(\mathrm{d}x_1, \ldots, \mathrm{d}x_k) \overset{\mathbb{P}}{\to} \int_{S^k} \prod_{i=1}^k f_i(x_i) \mu(\mathrm{d}x_1, \ldots, \mathrm{d}x_k)$$

for all $f_1, \ldots f_k \in \mathrm{BL}(S)$ and $k \in \mathbb{N}$. The following Lemma is essentially a version of Ethier & Kurtz (2005, Chapter 3 Proposition 4.6 b) extended to random measures.

LEMMA 2. *Let $\left(\mu_n^\omega\right)_{n\geqslant 1}$ be a sequence of random probability measures and $\mu$ a non-random probability measure on $S^\mathbb{N}$. Then $\mu_n^\omega \rightsquigarrow_{\mathbb{P}} \mu$ is equivalent to*

$$\int_{S^k} \prod_{i=1}^k f_i(x_i) \mu_n^\omega(\mathrm{d}x_1, \ldots, \mathrm{d}x_k) \overset{\mathbb{P}}{\to} \int_{S^k} \prod_{i=1}^k f_i(x_i) \mu(\mathrm{d}x_1, \ldots, \mathrm{d}x_k)$$

*for all $f_1, \ldots f_k \in \mathrm{BL}(S)$ and $k \in \mathbb{N}$.*

*Proof.* Suppose for any $k$ that the above convergence holds for all test functions $f_1, \ldots f_k \in \mathrm{BL}(S)$. We have shown in Lemma 1 that this is equivalent of convergence of the canonical projections $\mu_n^\omega \circ \pi_k^{-1}$ on $S^k$ (in probability) for any given $k$. Hence, using Proposition 1 for every subsequence $N \subset \mathbb{N}$ there is a subsequence $N' \subset N$ such that along $N'$

$$\mathbb{P}\left(\omega \in \Omega : \mu_n^{\omega} \circ \pi_k^{-1} \rightsquigarrow \mu \circ \pi_k^{-1} \text{ as } n \to \infty \quad \text{for all } k \in \mathbb{N}\right) = 1.$$

An application of Ethier & Kurtz (2005, Chapter 3 Proposition 4.6 b) concludes the proof. □

## S2. Proofs of Section 4

### S2·1. *Proofs for Section 4.1*

LEMMA 1. Given a random probability measure $\mu^{\omega}$ and random Markov kernel $K^{\omega}$, there exists an almost surely unique random probability measure $\mu^{\mathbb{N},\omega}$ on $S^{\mathbb{N}}$ such that

$$\mu^{\mathbb{N},\omega}(A_1 \times \ldots \times A_k \times E_{k+1}) = \int_{A_1} \mu^{\omega}(\mathrm{d}x_1) \int_{A_2} K^{\omega}(x_1, \mathrm{d}x_2) \ldots \int_{A_k} K^{\omega}(x_{k-1}, \mathrm{d}x_k)$$

for any $A_i \in \mathcal{B}(S)$ $(i = 1, \ldots, k)$, $k \in \mathbb{N}$ *and* $E_{k+1} = \times_{i=k+1}^{\infty} S$.

*Proof of Lemma* 1. For $\mathbb{P}-$almost all $\omega$, the existence and uniqueness of the distribution $\mu^{\mathbb{N},\omega}$ on $\{S^{\mathbb{N}}, \mathcal{B}(S)^{\mathbb{N}}\}$ can be obtained using the Ionescu-Tulcea extension theorem; see, e.g., Kallenberg (2006, Theorem 6.17) or Klenke (2013, Theorem 14.32). Measurability follows analogously by noting that $\omega \mapsto \mu^{\mathbb{N}}(\omega, A)$ is measurable for any $A \in \mathcal{E} = \{A_1 \times \ldots \times A_k \times E_{k+1};\, A_i \in \mathcal{B}(S), i = 1, \ldots, k, k \in \mathbb{N}\}$ and that $\mathcal{E}$ forms a $\pi-$system that generates $\mathcal{B}(S)^{\mathbb{N}}$. By Crauel (2003, Remark 3.2) this is enough to obtain measurability for every $A \in \mathcal{B}(S)^{\mathbb{N}}$. □

THEOREM 2. *If the following assumptions hold,*

*(T.1) the random probability measures $\left(\mu_n^{\omega}\right)_{n\geqslant 1}$ converge weakly in probability to a probability measure $\mu$ as $n \to \infty$,*

*(T.2) the random Markov transition kernels $\left(K_n^{\omega}\right)_{n\geqslant 1}$ satisfy*

$$\int \left|K_n^{\omega} f(x) - K f(x)\right| \mu_n^{\omega}(\mathrm{d}x) \to 0$$

*in probability as $n \to \infty$ for all $f \in \mathrm{BL}(S)$ where $K$ is a Markov transition kernel ,*

*(T.3) the transition kernel $K$ is such that $x \mapsto Kf(x)$* is continuous for any $f \in C_b(S)$,

*then, as $n \to \infty$, the measures $(\mu_n^{\mathbb{N},\omega})_{n\geqslant 1}$ on $S^{\mathbb{N}}$ converge weakly in probability to the measure $\mu^{\mathbb{N}}$ induced by the Markov chain with initial distribution $\mu$ and transition kernel $K$.*

*Proof of Theorem* 2. By Section S1·2 Lemma 2, we need to show that for any $k \geq 0$ and any $f_0, \ldots, f_k \in \mathrm{BL}(S)$

$$E^{\omega}\left\{f_0(X_{n,0}^{\omega}) \cdots f_k(X_{n,k}^{\omega})\right\} \xrightarrow{\mathbb{P}} E\left\{f_0(X_0) \cdots f_k(X_k)\right\} \tag{2.1}$$

where $E^{\omega}$, resp. $E$, denotes the expectation w.r.t. the law of $\mathbf{X}_n^{\omega}$, respectively w.r.t. the law of $\mathbf{X}$. We prove this by induction. For $k = 0$, this follows directly from $(T.1)$. Now assume that (2.1) is true for $k \geq 0$, i.e.

$$\left|E^{\omega}\left\{f_0(X_{n,0}^{\omega}) f_1(X_{n,1}^{\omega}) \cdots f_k(X_{n,k}^{\omega})\right\} - E\left\{f_0(X_0) f_1(X_1) \cdots f_k(X_k)\right\}\right| \xrightarrow{\mathbb{P}} 0.$$

By Lemma 1 this is equivalent to weak convergence in probability of the vector of the first $k$ states, i.e., for all $f \in C_b(S^k)$

$$E^{\omega}\left\{f(X_0^n, \ldots, X_k^n)\right\} \xrightarrow{\mathbb{P}} E\left\{f(X_0, \ldots, X_k)\right\}. \tag{2.2}$$

For $k+1$, we have

$$\left|E^{\omega}\left\{f_0(X_{n,0}^{\omega}) \cdots f_k(X_{n,k}^{\omega}) f_{k+1}(X_{n,k+1}^{\omega})\right\} - E\left\{f_0(X_0) \cdots f(X_k) f_{k+1}(X_{k+1})\right\}\right|$$

$$
\begin{aligned}
&= \left|E^{\omega}\left\{f_0(X^{\omega}_{n,0})\cdots f_k(X^{\omega}_{n,k})K^{\omega}_n f_{k+1}(X^{\omega}_{n,k})\right\} - E\left\{f_0(X_0)\cdots f(X_k)Kf_{k+1}(X_k)\right\}\right| \\
&\leq \left|E^{\omega}\left\{f_0(X^{\omega}_{n,0})\cdots f_k(X^{\omega}_{n,k})K^{\omega}_n f_{k+1}(X^{\omega}_{n,k}) - f_0(X^{\omega}_{n,0})\cdots f_k(X^{\omega}_{n,k})Kf_{k+1}(X^{\omega}_{n,k})\right\}\right| \\
&\qquad + \left|E^{\omega}\left\{f_0(X^{\omega}_{n,0})\cdots f_k(X^{\omega}_{n,k})Kf_{k+1}(X^{\omega}_{n,k})\right\} - E\left\{f_0(X_0)\cdots f_k(X_k)Kf_{k+1}(X_k)\right\}\right| \\
&\leq E^{\omega}\left\{\left|K^{\omega}_n f_{k+1}(X^{\omega}_{n,k}) - Kf_{k+1}(X^{\omega}_{n,k})\right|\right\} \qquad (2.3) \\
&\qquad + \left|E^{\omega}\left\{f_0(X^{\omega}_{n,0})\cdots f_k(X^{\omega}_{n,k})Kf_{k+1}(X^{\omega}_{n,k})\right\} - E\left\{f_0(X_0)\cdots f_k(X_k)Kf_{k+1}(X_k)\right\}\right|. \qquad (2.4)
\end{aligned}
$$

The term (2.3) converges due to (2). For the term (2.4), the function $Kf_{k+1}$ is bounded and it is assumed continuous so the function $f_0\cdots f_k Kf_{k+1} \in C_b(S^k)$. Hence this term vanishes by (2.2). □

### S2·2. *Some Auxiliary Results*

LEMMA 2. *Under Assumption* 1*, we have*

$$\varphi(\mathrm{d}\theta; \hat{\theta}^{\omega}_T, \Sigma/T) \rightsquigarrow_{\mathbb{P}^Y} \delta_{\bar{\theta}}(\mathrm{d}\theta)$$

*and*

$$\pi^{\omega}_T(\mathrm{d}\theta) \rightsquigarrow_{\mathbb{P}^Y} \delta_{\bar{\theta}}(\mathrm{d}\theta).$$

*Proof.* Using the moment generating function of the normal distribution, we have as $T \to \infty$

$$\int e^{u^{\mathrm{T}}\theta}\varphi(\theta; \hat{\theta}^{\omega}_T, \Sigma/T)\mathrm{d}\theta = \exp\left(u^{\mathrm{T}}\hat{\theta}^{\omega}_T + u^{\mathrm{T}}\Sigma u/2T\right) \xrightarrow{\mathbb{P}^Y} \exp(u^{\mathrm{T}}\bar{\theta}) = \int \exp(u^{\mathrm{T}}\theta)\delta_{\bar{\theta}}(\mathrm{d}\theta),$$

where $\delta_{\bar{\theta}}$ denotes the Dirac measure at $\bar{\theta}$ and thus $\varphi(\mathrm{d}\theta; \hat{\theta}^{\omega}_T, \Sigma/T) \rightsquigarrow_{\mathbb{P}^Y} \delta_{\bar{\theta}}(\mathrm{d}\theta)$ by Proposition 3. This implies that for $f \in C_b(\mathbb{R}^d)$

$$
\begin{aligned}
&\left|\int f(\theta)\pi^{\omega}_T(\theta)\mathrm{d}\theta - \int f(\theta)\delta_{\bar{\theta}}(\mathrm{d}\theta)\right| \\
&\leq \left|\int f(\theta)\pi^{\omega}_T(\theta)\mathrm{d}\theta - \int f(\theta)\varphi(\theta; \hat{\theta}^{\omega}_T, \Sigma/T)\mathrm{d}\theta\right| + \left|\int f(\theta)\varphi(\theta; \hat{\theta}^{\omega}_T, \Sigma/T)\mathrm{d}\theta - \int f(\theta)\delta_{\bar{\theta}}(\mathrm{d}\theta)\right| \\
&\leq \|f\|_{\infty}\int \left|\pi^{\omega}_T(\theta) - \varphi(\theta; \hat{\theta}^{\omega}_T, \Sigma/T)\right|\mathrm{d}\theta + \left|\int f(\theta)\varphi(\theta; \hat{\theta}^{\omega}_T, \Sigma/T)\mathrm{d}\theta - \int f(\theta)\delta_{\bar{\theta}}(\mathrm{d}\theta)\right|,
\end{aligned}
$$

where the first term on the r.h.s. converges to zero in probability under Assumption 1 while the second term converges to zero as $\varphi(\mathrm{d}\theta; \hat{\theta}^{\omega}_T, \Sigma/T) \rightsquigarrow_{\mathbb{P}^Y} \delta_{\bar{\theta}}(\mathrm{d}\theta)$. Hence, it follows that $\pi^{\omega}_T(\mathrm{d}\theta) \rightsquigarrow_{\mathbb{P}^Y} \delta_{\bar{\theta}}(\mathrm{d}\theta)$. □

To analyse the asymptotic properties of the pseudo-marginal algorithm, we rescale the parameter component. A simple change of variables and the fact that convergence in total variation in probability implies weak convergence in probability shows that the following result holds.

LEMMA 3. *Under Assumption* 1*, we have*

$$\int \left|\tilde{\pi}^{\omega}_T(\tilde{\theta}) - \varphi(\tilde{\theta}; 0, \Sigma)\right|\mathrm{d}\theta \xrightarrow{\mathbb{P}^Y} 0, \quad \textit{as} \quad T \to \infty,$$

*and thus* $\tilde{\pi}^{\omega}_T(\mathrm{d}\tilde{\theta}) \rightsquigarrow_{\mathbb{P}^Y} \varphi(\mathrm{d}\tilde{\theta}; 0, \Sigma)$.

LEMMA 4 (CONVERGENCE OF MARGINAL DISTRIBUTIONS). *Under Assumptions* 1 *and* 2*, the marginal distribution of the proposal at stationarity*

$$\pi^{\omega}_T q_T(\mathrm{d}\vartheta) = \int \pi^{\omega}_T(\mathrm{d}\theta)q_T(\theta, \mathrm{d}\vartheta)$$

*satisfies*

$$\pi^{\omega}_T q_T(\mathrm{d}\vartheta) \rightsquigarrow_{\mathbb{P}^Y} \delta_{\bar{\theta}}(\mathrm{d}\vartheta).$$

*Proof.* Let $f \in \mathrm{BL}(\mathbb{R})$, then we have

$$\begin{aligned}\left|\int f(\vartheta)\pi_T^{\omega} q_T(\mathrm{d}\vartheta) - f(\bar\theta)\right| &= \left|\int f(\theta+\xi/\sqrt{T})\int \pi_T^{\omega}(\mathrm{d}\theta)\nu(\mathrm{d}\xi) - f(\bar\theta)\right| \\ &\le \left|\iint \left(f(\theta+\xi/\sqrt{T}) - f(\theta)\right)\nu(\mathrm{d}\xi)\pi_T^{\omega}(\mathrm{d}\theta)\right| + \left|\iint f(\theta)\pi_T^{\omega}(\mathrm{d}\theta)\nu(\mathrm{d}\xi) - f(\bar\theta)\right| \\ &\le \iint \left|f(\theta+\xi/\sqrt{T}) - f(\theta)\right|\nu(\mathrm{d}\xi)\pi_T^{\omega}(\mathrm{d}\theta) + \left|\int f(\theta)\pi_T^{\omega}(\mathrm{d}\theta) - f(\bar\theta)\right|.\end{aligned}$$

The second term on the r.h.s. vanishes due to Lemma 2. For the first term we use the fact that $f$ is bounded Lipschitz, hence

$$\begin{aligned}\iint \left|f(\theta+\xi/\sqrt{T}) - f(\theta)\right|\nu(\mathrm{d}\xi)\pi_T^{\omega}(\mathrm{d}\theta) &\le \|f\|_{\mathrm{BL}}\iint \min\left\{1, \frac{\|\xi\|}{\sqrt{T}}\right\}\nu(\mathrm{d}\xi)\pi_T^{\omega}(\mathrm{d}\theta) \\ &= \|f\|_{\mathrm{BL}}\iint \min\left\{1, \frac{\|\xi\|}{\sqrt{T}}\right\}\nu(\mathrm{d}\xi) \to 0.\end{aligned}$$

The proof of the following Lemmas are straightforward and thus omitted.

LEMMA 5. *The map* $x \mapsto \min(1, ae^x)$ *with* $a > 0$ *is* $1-$*Lipschitz, i.e., for all* $x, y \in \mathbb{R}$

$$\left|\min\left(1, ae^x\right) - \min\left(1, ae^y\right)\right| \le |x-y|.$$

LEMMA 6. *Under Assumption* 3

*(i) the function*

$$\theta \mapsto d_{\mathrm{BL}}\left[\varphi\left\{\cdot\,; \sigma^2(\theta)/2, \sigma^2(\theta)\right)\}, \varphi\left\{\cdot\,; \sigma^2(\bar\theta)/2, \sigma^2(\bar\theta)\right\} \mid \mathcal{Y}_T\right]$$

*is bounded for all* $\theta$ *and continuous at* $\bar\theta$*;*

*(ii) for all* $f \in \mathrm{BL}(\mathbb{R})$ *the functions*

$$\theta \mapsto \left|\int f(z)\varphi\left\{\mathrm{d}z; \sigma^2(\theta)/2, \sigma^2(\theta)\right\} - \int f(z)\varphi\left\{\mathrm{d}z; \sigma^2(\bar\theta)/2, \sigma^2(\bar\theta)\right\}\right|$$

*are bounded for all* $\theta$ *and continuous at* $\bar\theta$.

### S2·3. *Proof of Theorem* 1

In order to prove Theorem 1, we need to prove Propositions 1, 2 and 3 of Section 4·3.

PROPOSITION 1. *Under Assumptions* 1 *and* 3, *we have*

$$\tilde\pi_T^{\omega}(\mathrm{d}\tilde\theta, \mathrm{d}z) \to \tilde\pi(\mathrm{d}\tilde\theta, \mathrm{d}z),$$

*weakly in* $\mathbb{P}^Y$*-probability as* $T \to \infty$ *where* $\tilde\pi_T^{\omega}(\mathrm{d}\tilde\theta, \mathrm{d}z) = \tilde\pi_T^{\omega}(\mathrm{d}\tilde\theta)\exp(z)\,\tilde g_T^{\omega}(\mathrm{d}z \mid \tilde\theta)$.

*Proof of Proposition* 1. As established in Lemma 1, it is enough to check convergence for products of bounded Lipschitz functions. Now, without loss of generality, assume that $\|f_1\|_\infty, \|f_2\|_\infty \le 1/2$. Then we have

$$\begin{aligned}&\left|\iint f_1(\tilde\theta)f_2(z)\tilde\pi_T^{\omega}(\mathrm{d}\tilde\theta)e^z\tilde g_T^{\omega}(\mathrm{d}z \mid \tilde\theta) - \iint f_1(\tilde\theta)f_2(z)\varphi(\mathrm{d}\tilde\theta; 0, \Sigma)\varphi\left\{\mathrm{d}z; \sigma^2(\bar\theta)/2, \sigma^2(\bar\theta)\right\}\right| \\ &\le \iint e^z\tilde g_T^{\omega}(z \mid \tilde\theta)\mathrm{d}z\left|\tilde\pi_T^{\omega}(\tilde\theta) - \varphi(\tilde\theta; 0, \Sigma)\right|\mathrm{d}\tilde\theta \\ &\quad + \int \varphi(\tilde\theta; 0, \Sigma)\left|\int f_2(z)e^z\tilde g_T^{\omega}(\mathrm{d}z \mid \tilde\theta) - \int f_2(z)\varphi\left\{\mathrm{d}z; \sigma^2(\bar\theta)/2, \sigma^2(\bar\theta)\right\}\right|\mathrm{d}\tilde\theta \\ &\le \int\left|\tilde\pi_T^{\omega}(\mathrm{d}\tilde\theta) - \varphi(\tilde\theta; 0, \Sigma)\right|\mathrm{d}\tilde\theta \qquad (2.5)\end{aligned}$$

$$+\int \varphi(\theta;\hat{\theta}_T^{\omega},\Sigma/T)\left|\int f_2(z)e^z g_T^{\omega}(\mathrm{d}z\mid\theta)-\int f_2(z)\varphi\left\{\mathrm{d}z;\sigma^2(\theta)2,\sigma^2(\theta)\right\}\right|\mathrm{d}\theta \tag{2.6}$$

$$+\int \varphi(\theta;\hat{\theta}_T^{\omega},\Sigma/T)\left|\int f_2(z)\varphi\left\{\mathrm{d}z;\sigma^2(\theta)/2,\sigma^2(\theta)\right\}-\int f_2(z)\varphi\left\{\mathrm{d}z;\sigma^2(\bar{\theta})/2,\sigma^2(\bar{\theta})\right\}\right|\mathrm{d}\theta \tag{2.7}$$

The term (2.5) converges to zero in $\mathbb{P}^Y$-probability by Lemma 3. For (2.6), write $B(\bar{\theta})\subset\Theta$ for the $\varepsilon$-ball on which the uniform CLT in Assumption 3 holds, that is

$$\sup_{\theta\in B(\bar{\theta})} h_T(\theta)=\sup_{\theta\in B(\bar{\theta})}\left|\int f_2(z)e^z g_T^{\omega}(\mathrm{d}z\mid\theta)\mathrm{d}z-\int f_2(z)\varphi\left\{\mathrm{d}z;\sigma^2(\theta)/2,\sigma^2(\theta)\right\}\mathrm{d}z\right|\xrightarrow{\mathbb{P}^Y}0.$$

We can bound (2.6) as follows

$$\begin{aligned}&\int_{B(\bar{\theta})}\varphi(\theta;\hat{\theta}_T^{\omega},\Sigma/T)\left|\int f_2(z)e^z g_T^{\omega}(\mathrm{d}z\mid\theta)-\int f_2(z)\varphi\left\{\mathrm{d}z;\sigma^2(\theta)/2,\sigma^2(\theta)\right\}\right|\mathrm{d}\theta\\&+\int_{B(\bar{\theta})^{\complement}}\varphi(\theta;\hat{\theta}_T^{\omega},\Sigma/T)\left|\int f_2(z)e^z g_T^{\omega}(\mathrm{d}z\mid\theta)-\int f_2(z)\varphi\left\{\mathrm{d}z;\sigma^2(\theta)/2,\sigma^2(\theta)\right\}\right|\mathrm{d}\theta\\&\le\sup_{\theta\in B(\bar{\theta})}h_T(\theta)+\int_{B(\bar{\theta})^{\complement}}\varphi(\theta;\hat{\theta}_T^{\omega},\Sigma/T)\mathrm{d}\theta,\end{aligned}$$

since $\|f_2\|_\infty\le 1/2$. We have already mentioned that the first term vanishes in probability whereas for the second term we have

$$\int_{B(\bar{\theta})^{\complement}}\varphi(\theta;\hat{\theta}_T^{\omega},\Sigma/T)\mathrm{d}\theta\xrightarrow{\mathbb{P}^Y}\delta_{\bar{\theta}}\{B(\bar{\theta})^{\complement}\}=0,$$

by Lemma 2. Thus (2.6) vanishes in $\mathbb{P}^Y$-probability. Finally we consider (2.7). By Lemma 6

$$h(\theta)=\left|\int f_2(z)\varphi\left\{\mathrm{d}z;\sigma^2(\theta)/2,\sigma^2(\theta)\right\}-\int f_2(z)\varphi\left\{\mathrm{d}z;\sigma^2(\bar{\theta})/2,\sigma^2(\bar{\theta})\right\}\right|$$

is bounded and continuous at $\bar{\theta}$. Since $\varphi(\mathrm{d}\theta;\hat{\theta}_T^{\omega},\Sigma/T)$ converges weakly in probability to a point mass in $\bar{\theta}$ (by Lemma 2) we can conclude that

$$\int f(\theta)\varphi(\mathrm{d}\theta;\hat{\theta}_T^{\omega},\Sigma/T)\xrightarrow{\mathbb{P}^Y}\int f(\theta)\delta_{\bar{\theta}}(\mathrm{d}\theta)$$

for every bounded function $f$ which is continuous at $\overline{\theta}$. In particular,

$$\int h(\theta)\varphi(\mathrm{d}\theta;\hat{\theta}_T^{\omega},\Sigma/T)\xrightarrow{\mathbb{P}^Y}0.$$

PROPOSITION 2. *Under Assumptions* 1, 2 *and* 3, *as* $T\to\infty$ *we have for any* $f\in\mathrm{BL}(\mathbb{R}^{d+1})$

$$\int|\tilde{P}_T^{\omega}f(\theta,z)-\tilde{P}f(\theta,z)|\tilde{\pi}_T^{\omega}(\mathrm{d}\theta,\mathrm{d}z)\to 0,\quad \textit{in } \mathbb{P}^Y\textit{-probability}.$$

*Proof of Proposition* 2. Let $f\in\mathrm{BL}(\mathbb{R}^{d+1})$. Denote

$$\Pi_T^{\omega}f(\tilde{\theta},z)=\iint f(\tilde{\theta}',z')\tilde{\alpha}_T^{\omega}\{(\tilde{\theta},z),(\tilde{\theta}',z')\}\tilde{q}(\tilde{\theta},\mathrm{d}\tilde{\theta}')\tilde{g}_T^{\omega}(\mathrm{d}z'\mid\tilde{\theta}')$$

and

$$\Pi f(\tilde{\theta},z)=\iint f(\tilde{\theta}',z')\tilde{\alpha}\{(\tilde{\theta},z),(\tilde{\theta}',z')\}\tilde{q}(\tilde{\theta},\mathrm{d}\tilde{\theta}')g(\mathrm{d}z'\mid\overline{\theta}),$$

where $g(\cdot \mid \vartheta) = \varphi\{\cdot\,; -\sigma^2(\vartheta)/2, \sigma^2(\vartheta)\}$. Then we have

$$\tilde{P}_T^{\omega} f(\tilde{\theta}, z) = \Pi_T^{\omega} f(\tilde{\theta}, z) + f(\tilde{\theta}, z)\left\{1 - \Pi_T^{\omega} 1(\tilde{\theta}, z)\right\}$$

and

$$\tilde{P} f(\tilde{\theta}, z) = \Pi f(\tilde{\theta}, z) + f(\tilde{\theta}, z)\left\{1 - \Pi 1(\tilde{\theta}, z)\right\}. \tag{2.8}$$

Because

$$\begin{aligned}
&E^{\omega}\left\{\left|\tilde{P}_T^{\omega} f(\tilde{\vartheta}_0^T, Z_0^T) - \tilde{P} f(\tilde{\vartheta}_0^T, Z_0^T)\right|\right\} \\
&= E^{\omega}\Big[\Big|\Pi_T^{\omega} f(\tilde{\vartheta}_0^T, Z_0^T) + f(\tilde{\vartheta}_0^T, Z_0^T)\left\{1 - \Pi_T^{\omega} 1(\tilde{\vartheta}_0^T, Z_0^T)\right\} \\
&\qquad - \Pi f(\tilde{\vartheta}_0^T, Z_0^T) - f(\tilde{\vartheta}_0^T, Z_0^T)\left\{1 - \Pi 1(\tilde{\vartheta}_0^T, Z_0^T)\right\}\Big|\Big] \\
&\leq E^{\omega}\left\{\left|\Pi_T^{\omega} f(\tilde{\vartheta}_0^T, Z_0^T) - \Pi f(\tilde{\vartheta}_0^T, Z_0^T)\right|\right\} + E^{\omega}\left\{\left|\Pi_T^{\omega} 1(\tilde{\vartheta}_0^T, Z_0^T) - \Pi 1(\tilde{\vartheta}_0^T, Z_0^T)\right|\right\}
\end{aligned}$$

and $1 \in \mathrm{BL}(\mathbb{R}^{d+1})$ it is sufficient to show that for any choice of $f \in \mathrm{BL}(\mathbb{R}^{d+1})$ we have $E^{\omega}\left\{\left|\Pi_T^{\omega} f(\tilde{\theta}, z) - \Pi f(\tilde{\theta}, z)\right|\right\} \xrightarrow{\mathbb{P}^Y} 0$. Thus

$$\begin{aligned}
&E^{\omega}\left\{\left|\Pi_T^{\omega} f(\tilde{\theta}, z) - \Pi f(\tilde{\theta}, z)\right|\right\} \\
&= \iint \tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}, \mathrm{d}z)\bigg|\iint \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}')\tilde{\alpha}_T^{\omega}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}') \\
&\qquad - \iint \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}')\tilde{\alpha}\{(\tilde{\theta}, z), (\tilde{\theta}', z')\} f(\tilde{\theta}', z') g(\mathrm{d}z' \mid \overline{\theta})\bigg| \\
&= \iint e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\bigg|\iint \min\left\{\tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \tilde{\pi}_T^{\omega}(\tilde{\theta}')\tilde{q}(\tilde{\theta}', \tilde{\theta})e^{z'-z}\right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}' \\
&\qquad - \iint \tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}')\tilde{\alpha}\{(\tilde{\theta}, z), (\tilde{\theta}', z')\} f(\tilde{\theta}', z') g(\mathrm{d}z' \mid \overline{\theta})\mathrm{d}\tilde{\theta}'\bigg|\mathrm{d}\tilde{\theta} \\
&\leq \iint e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\bigg|\iint \min\left\{\tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \tilde{\pi}_T^{\omega}(\tilde{\theta}')\tilde{q}(\tilde{\theta}', \tilde{\theta})e^{z'-z}\right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}' \\
&\qquad - \iint \min\left\{\varphi(\tilde{\theta}; 0, \Sigma)\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \varphi(\tilde{\theta}'; 0, \Sigma)\tilde{q}(\tilde{\theta}', \tilde{\theta})e^{z'-z}\right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}'\bigg|\mathrm{d}\tilde{\theta} \\
&+ \iint e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\bigg|\iint \min\left\{\varphi(\tilde{\theta}; 0, \Sigma)\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \varphi(\tilde{\theta}'; 0, \Sigma)\tilde{q}(\tilde{\theta}', \tilde{\theta})e^{z'-z}\right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}' \\
&\qquad - \iint \tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}')\tilde{\alpha}\{(\tilde{\theta}, z), (\tilde{\theta}', z')\} f(\tilde{\theta}', z') g(\mathrm{d}z' \mid \overline{\theta})\mathrm{d}\tilde{\theta}'\bigg|\mathrm{d}\tilde{\theta}.
\end{aligned} \tag{2.9}$$

By taking $\varphi(\tilde{\theta}; 0, \Sigma)$ out in last two lines of (2.9), this can be rewritten as

$$\begin{aligned}
&\iint e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\bigg|\iint \min\left\{\tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \tilde{\pi}_T^{\omega}(\tilde{\theta}')\tilde{q}(\tilde{\theta}', \tilde{\theta})e^{z'-z}\right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}' \\
&\qquad - \iint \min\left\{\varphi(\tilde{\theta}; 0, \Sigma)\tilde{q}(\tilde{\theta}, \tilde{\theta}'), \varphi(\tilde{\theta}'; 0, \Sigma)\tilde{q}(\tilde{\theta}', \tilde{\theta})e^{z'-z}\right\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}'\bigg|\mathrm{d}\tilde{\theta} \qquad (2.10) \\
&+ \iint e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\bigg|\iint \varphi(\tilde{\theta}; 0, \Sigma)\tilde{q}(\tilde{\theta}, \tilde{\theta}')\tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} f(\tilde{\theta}', z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}'
\end{aligned}$$

$$- \iint \tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta},\tilde{\theta}')\tilde{\alpha}\big\{(\tilde{\theta},z),(\tilde{\theta}',z')\big\} f(\tilde{\theta}',z') g(\mathrm{d}z' \mid \overline{\theta})\mathrm{d}\tilde{\theta}'\bigg|\mathrm{d}\tilde{\theta}. \tag{2.11}$$

For (2.10), we use the inequality $|\min(a,b) - \min(c,d)| \leq |a-c| + |b-d|$:

$$\begin{aligned}
&\iint e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\bigg| \iint \min\Big\{\tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta},\tilde{\theta}'), \tilde{\pi}_T^{\omega}(\tilde{\theta}')\tilde{q}(\tilde{\theta}',\tilde{\theta})e^{z'-z}\Big\} f(\tilde{\theta}',z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}') \\
&\qquad - \min\Big\{\varphi(\tilde{\theta};0,\Sigma)\tilde{q}(\tilde{\theta},\tilde{\theta}'), \varphi(\tilde{\theta}';0,\Sigma)\tilde{q}(\tilde{\theta}',\tilde{\theta})e^{z'-z}\Big\} f(\tilde{\theta}',z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}'\bigg|\mathrm{d}\tilde{\theta} \\
&\leq \|f\|_{\infty} \iiiint e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\tilde{q}(\tilde{\theta},\mathrm{d}\tilde{\theta}')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}') \left|\tilde{\pi}_T^{\omega}(\tilde{\theta}) - \varphi(\tilde{\theta};0,\Sigma)\right| \mathrm{d}\tilde{\theta} \\
&\qquad + \|f\|_{\infty} \iiiint \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})e^{z'}\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\tilde{q}(\tilde{\theta}',\tilde{\theta}) \left|\tilde{\pi}_T^{\omega}(\tilde{\theta}') - \varphi(\tilde{\theta}';0,\Sigma)\right| \mathrm{d}\tilde{\theta}'\mathrm{d}\tilde{\theta} \\
&= 2\|f\|_{\infty} \int \left|\tilde{\pi}_T^{\omega}(\tilde{\theta}) - \varphi(\tilde{\theta};0,\Sigma)\right| \mathrm{d}\tilde{\theta} \xrightarrow{\mathbb{P}^Y} 0,
\end{aligned}$$

by Lemma 3. For the part (2.11) note that

$$\begin{aligned}
&\iint e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\bigg| \iint \varphi(\tilde{\theta};0,\Sigma)\tilde{q}(\tilde{\theta},\mathrm{d}\tilde{\theta}')\tilde{\alpha}\big\{(\tilde{\theta},z),(\tilde{\theta}',z')\big\} f(\tilde{\theta}',z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}') \\
&\qquad - \iint \tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta},\mathrm{d}\tilde{\theta}')\tilde{\alpha}\big\{(\tilde{\theta},z),(\tilde{\theta}',z')\big\} f(\tilde{\theta}',z') g(\mathrm{d}z' \mid \overline{\theta})\bigg|\mathrm{d}\tilde{\theta} \\
&\leq \iint e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\bigg| \iint \varphi(\tilde{\theta};0,\Sigma)\tilde{q}(\tilde{\theta},\mathrm{d}\tilde{\theta}')\tilde{\alpha}\big\{(\tilde{\theta},z),(\tilde{\theta}',z')\big\} f(\tilde{\theta}',z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}') \\
&\qquad - \iint \tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta},\mathrm{d}\tilde{\theta}')\tilde{\alpha}\big\{(\tilde{\theta},z),(\tilde{\theta}',z')\big\} f(\tilde{\theta}',z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\bigg|\mathrm{d}\tilde{\theta} &(2.12) \\
&\qquad + \iint \tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta})e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\bigg| \iint \tilde{q}(\tilde{\theta},\mathrm{d}\tilde{\theta}')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\tilde{\alpha}\big\{(\tilde{\theta},z),(\tilde{\theta}',z')\big\} f(\tilde{\theta}',z') \\
&\qquad - \iint \tilde{q}(\tilde{\theta},\mathrm{d}\tilde{\theta}') g(\mathrm{d}z' \mid \overline{\theta})\tilde{\alpha}\big\{(\tilde{\theta},z),(\tilde{\theta}',z')\big\} f(\tilde{\theta}',z')\bigg|. &(2.13)
\end{aligned}$$

For the first part (2.12) we have

$$\begin{aligned}
&\iint e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\bigg| \iint \varphi(\tilde{\theta};0,\Sigma)\tilde{q}(\tilde{\theta},\tilde{\theta}')\tilde{\alpha}\big\{(\tilde{\theta},z),(\tilde{\theta}',z')\big\} f(\tilde{\theta}',z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}' \\
&\qquad - \iint \tilde{\pi}_T^{\omega}(\tilde{\theta})\tilde{q}(\tilde{\theta},\tilde{\theta}')\tilde{\alpha}\big\{(\tilde{\theta},z),(\tilde{\theta}',z')\big\} f(\tilde{\theta}',z')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\mathrm{d}\tilde{\theta}'\bigg|\mathrm{d}\tilde{\theta} \\
&\leq \|f\|_{\infty} \iiiint e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\tilde{q}(\tilde{\theta},\mathrm{d}\tilde{\theta}')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}') \left|\varphi(\tilde{\theta};0,\Sigma) - \tilde{\pi}_T^{\omega}(\tilde{\theta})\right| \mathrm{d}\tilde{\theta} \\
&= \|f\|_{\infty} \int \left|\varphi(\tilde{\theta};0,\Sigma) - \tilde{\pi}_T^{\omega}(\tilde{\theta})\right| \mathrm{d}\tilde{\theta} \xrightarrow{\mathbb{P}^Y} 0,
\end{aligned}$$

again by Lemma 3. The second part (2.13)

$$\begin{aligned}
&\iint \tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta})e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\bigg| \iint \tilde{q}(\tilde{\theta},\mathrm{d}\tilde{\theta}')\tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\tilde{\alpha}\big\{(\tilde{\theta},z),(\tilde{\theta}',z')\big\} f(\tilde{\theta}',z') \\
&\qquad - \iint \tilde{q}(\tilde{\theta},\mathrm{d}\tilde{\theta}') g(\mathrm{d}z' \mid \overline{\theta})\tilde{\alpha}\big\{(\tilde{\theta},z),(\tilde{\theta}',z')\big\} f(\tilde{\theta}',z')\bigg| \\
&\leq \iiint \tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta})e^{z}\tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta})\tilde{q}(\tilde{\theta},\mathrm{d}\tilde{\theta}')\bigg| \int \tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}')\tilde{\alpha}\big\{(\tilde{\theta},z),(\tilde{\theta}',z')\big\} f(\tilde{\theta}',z')
\end{aligned}$$

$$- \int \tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} g(\mathrm{d}z' \mid \hat{\theta}_T^{\omega} + \tilde{\theta}'/\sqrt{T}) f(\tilde{\theta}', z') \Big| \tag{2.14}$$

$$+ \iiint \tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}) e^{z} \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta}) \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') \Big| \int g(\mathrm{d}z' \mid \hat{\theta}_T^{\omega} + \tilde{\theta}'/\sqrt{T}) f(\tilde{\theta}', z') \tilde{\alpha}\big\{(\tilde{\theta}, z), (\tilde{\theta}', z')\big\} f(\tilde{\theta}', z')$$

$$- \int \tilde{\alpha}\{(\tilde{\theta}, z), (\tilde{\theta}', z')\} g(\mathrm{d}z' \mid \bar{\theta}) f(\tilde{\theta}', z') \Big| \tag{2.15}$$

We first consider (2.14) using $\theta = \hat{\theta}_T^{\omega} + \tilde{\theta}/\sqrt{T}$, and similarly for $\theta'$,

$$\iiint \tilde{\pi}_T^{\omega}(\mathrm{d}\tilde{\theta}) e^{z} \tilde{g}_T^{\omega}(\mathrm{d}z \mid \tilde{\theta}) \tilde{q}(\tilde{\theta}, \mathrm{d}\tilde{\theta}') \Big| \int \min\left\{1, \frac{\varphi(\tilde{\theta}'; 0, \Sigma)}{\varphi(\tilde{\theta}; 0, \Sigma)} \frac{\tilde{q}(\tilde{\theta}', \tilde{\theta})}{\tilde{q}(\tilde{\theta}, \tilde{\theta}')} e^{z'-z}\right\} \tilde{g}_T^{\omega}(\mathrm{d}z' \mid \tilde{\theta}') f(\tilde{\theta}', z')$$

$$- \int \min\left\{1, \frac{\varphi(\tilde{\theta}'; 0, \Sigma)}{\varphi(\tilde{\theta}; 0, \Sigma)} \frac{\tilde{q}(\tilde{\theta}', \tilde{\theta})}{\tilde{q}(\tilde{\theta}, \tilde{\theta}')} e^{z'-z}\right\} g(\mathrm{d}z' \mid \hat{\theta}_T^{\omega} + \tilde{\theta}'/\sqrt{T}) f(\tilde{\theta}', z') \Big|$$

$$= \iiint \pi_T^{\omega}(\mathrm{d}\theta) e^{z} g_T^{\omega}(\mathrm{d}z \mid \theta) q_T(\theta, \mathrm{d}\theta') \Big| \int \min\left\{1, \frac{\varphi(\theta'; \hat{\theta}_T^{\omega}, \Sigma/T)}{\varphi(\theta; \hat{\theta}_T^{\omega}, \Sigma/T)} \frac{q_T(\theta', \theta)}{q_T(\theta, \theta')} e^{z'-z}\right\} g_T^{\omega}(\mathrm{d}z' \mid \theta') f\big\{\sqrt{T}(\theta' - \hat{\theta}_T^{\omega}), z'\big\}$$

$$- \int \min\left\{1, \frac{\varphi(\theta'; \hat{\theta}_T^{\omega}, \Sigma/T)}{\varphi(\theta; \hat{\theta}_T^{\omega}, \Sigma/T)} \frac{q_T(\theta', \theta)}{q_T(\theta, \theta')} e^{z'-z}\right\} g(\mathrm{d}z' \mid \theta') f\big\{\sqrt{T}(\theta' - \hat{\theta}_T^{\omega}), z'\big\} \Big|$$

In the rest of the proof, without loss of generality, we will consider $f$ such that $\|f\|_{\mathrm{L}} \le 1$

$$\begin{aligned} &\Big| f\big\{\sqrt{T}(\theta' - \hat{\theta}_T^{\omega}), x\big\} - f\big\{\sqrt{T}(\theta' - \hat{\theta}_T^{\omega}), y\big\} \Big| \\ &\le d\left[\big\{\sqrt{T}(\theta' - \hat{\theta}_T^{\omega}), x\big\}, \big\{\sqrt{T}(\theta' - \hat{\theta}_T^{\omega}), y\big\}\right] = |x - y| \end{aligned}$$

and thus $x \mapsto f\big\{\sqrt{T}(\theta' - \hat{\theta}_T^{\omega}), x\big\}$ is Lipschitz with coefficient 1 uniformly in $T$. Moreover, due to Lemma 5, the map

$$z' \mapsto \min\left\{1, e^{-z} \frac{\varphi(\theta'; \hat{\theta}_T^{\omega}, \Sigma/T)}{\varphi(\theta; \hat{\theta}_T^{\omega}, \Sigma/T)} \frac{q_T(\theta', \theta)}{q_T(\theta, \theta')} e^{z'}\right\}$$

is Lipschitz with Lipschitz constant 1 uniformly for all $\theta, \theta', z$ and $T$. Thus, using the triangle inequality, we can write

$$\iiint \pi_T^{\omega}(\mathrm{d}\theta) e^{z} g_T^{\omega}(\mathrm{d}z \mid \theta) q_T(\theta, \mathrm{d}\theta') \Big| \int \min\left\{1, \frac{\varphi(\theta'; \hat{\theta}_T^{\omega}, \Sigma/T)}{\varphi(\theta; \hat{\theta}_T^{\omega}, \Sigma/T)} e^{z'-z}\right\} f\big\{\sqrt{T}(\theta' - \hat{\theta}_T^{\omega}), z'\big\} g_T^{\omega}(z' \mid \theta')$$

$$- \min\left\{1, \frac{\varphi(\theta'; \hat{\theta}_T^{\omega}, \Sigma/T)}{\varphi(\theta; \hat{\theta}_T^{\omega}, \Sigma/T)} e^{z'-z}\right\} f\big\{\sqrt{T}(\theta' - \hat{\theta}_T^{\omega}), z'\big\} g(z' \mid \theta') \Big| \mathrm{d}z'$$

$$\le 2 \iint \pi_T^{\omega}(\mathrm{d}\theta) q_T(\theta, \mathrm{d}\theta') \cdot \sup_{f \in \mathrm{BL}(\mathbb{R}),\, \|f\|_{\mathrm{BL}} \le 1} \Big| \int f(z') g_T^{\omega}(\mathrm{d}z' \mid \theta') - \int f(z') g(\mathrm{d}z' \mid \theta') \Big| \mathrm{d}\theta$$

$$= 2 \iint \pi_T^{\omega}(\mathrm{d}\theta) q_T(\theta, \mathrm{d}\theta') d_{\mathrm{BL}}\big\{g_T^{\omega}(\cdot|\theta'), g(\cdot|\theta')\big\}$$

$$= 2 \int_{B(\bar{\theta})} \pi_T^{\omega} q_T(\mathrm{d}\theta') d_{\mathrm{BL}}\big\{g_T^{\omega}(\cdot|\theta'), g(\cdot|\theta')\big\} + 2 \int_{B(\bar{\theta})^{\complement}} \pi_T^{\omega} q_T(\mathrm{d}\theta') d_{\mathrm{BL}}\left(g_T^{\omega}(\cdot|\theta'), g(\cdot|\theta')\right),$$

where $B(\bar{\theta})$ is given in Assumption 3. Since the bounded Lipschitz norm metrizes weak convergence (for non-random probability measures) we know that for $\theta' \in B(\bar{\theta})$

$$d_{\mathrm{BL}}\left(g_T^{\omega}(\cdot|\theta'), g(\cdot|\theta')\right) = \sup_{f \in \mathrm{BL}(\mathbb{R}),\, \|f\|_{\mathrm{BL}} \le 1} \Big| \int f(z') g_T^{\omega}(\mathrm{d}z' \mid \theta') - \int f(z') g(\mathrm{d}z' \mid \theta') \Big|$$

vanishes in $\mathbb{P}^Y$-probability by Assumption 3. From Lemma 4 we know that the marginal distribution of the proposal at stationarity $\pi_T^\omega q_T(\mathrm{d}\theta') = \int \pi_T^\omega(\mathrm{d}\theta) q(\theta, \mathrm{d}\theta')$ concentrates around the true parameter value. Since the bounded Lipschitz metric cannot exceed 1 we have

$$\int \pi_T^\omega q_T(\mathrm{d}\theta') \mathbb{I}_{B(\bar\theta)^\complement}(\theta') d_{\mathrm{BL}}\left(g_T^\omega(\cdot|\theta'), g(\cdot|\theta')\right) \leq \pi_T^\omega q_T\{B(\bar\theta)^\complement\} \xrightarrow{\mathbb{P}^Y} \delta_{\bar\theta}\{B(\bar\theta)^\complement\} = 0.$$

In addition from Assumption 3

$$\left| \int_{B(\bar\theta)} \pi_T^\omega q_T(\mathrm{d}\theta') d_{\mathrm{BL}}\{g_T^\omega(\cdot|\theta'), g(\cdot|\theta')\} \right| \leq \sup_{\theta \in B(\bar\theta)} \left| d_{\mathrm{BL}}\{g_T^\omega(\cdot|\theta), g(\cdot|\theta)\} \right| \xrightarrow{\mathbb{P}^Y} 0.$$

Finally, using a similar argument for (2.15) we have

$$\begin{aligned}
&\iiint \tilde\pi_T^\omega(\mathrm{d}\tilde\theta) e^z \tilde g_T^\omega(\mathrm{d}z \mid \tilde\theta) \tilde q(\tilde\theta, \mathrm{d}\tilde\theta') \left| \int g(z' \mid \hat\theta_T^\omega + \tilde\theta'/\sqrt{T}) f(\tilde\theta', z') \tilde\alpha\{(\tilde\theta, z), (\tilde\theta', z')\} f(\tilde\theta', z') \right. \\
&\qquad \left. - \tilde q(\tilde\theta, \mathrm{d}\tilde\theta') \tilde\alpha\{(\tilde\theta, z), (\tilde\theta', z')\} g(z' \mid \bar\theta) f(\tilde\theta', z') \right| \mathrm{d}z' \\
&\leq 2 \iint \pi_T^\omega(\theta) q_T(\theta, \theta') \mathrm{d}\theta d_{\mathrm{BL}}\left(g(\cdot|\theta'), g(\cdot|\bar\theta)\right) \mathrm{d}\theta'. \qquad (2.16)
\end{aligned}$$

By Lemma 6 the bounded Lipschitz metric, $d_{\mathrm{BL}}\{g(\cdot|\theta'), g(\cdot|\bar\theta)\}$, is bounded and continuous at $\bar\theta$. Thus (2.16) converges to zero by Lemma 4. □

PROPOSITION 3. *Under Assumption 2, the map $(\theta, z) \mapsto \tilde P f(\theta, z)$ is continuous for every $f \in C_b(\mathbb{R}^{d+1})$.*

*Proof of Proposition* 3. Without loss of generality let $\|f\|_\infty \leq 1$, consider $(\theta^*, z^*) \in \Theta \times \mathbb{R}$ and denote $(\theta_n, z_n)_{n \in \mathbb{N}}$ a sequence converging to $(\theta^*, z^*)$ as $n \to \infty$. Using the decomposition (2.8) we have

$$\begin{aligned}
&\left| \tilde P f(\theta_n, z_n) - \tilde P f(\theta^*, z^*) \right| \\
&= \left| \Pi f(\theta_n, z_n) + f(\theta_n, z_n)\{1 - \Pi 1(\theta_n, z_n)\} - \Pi f(\theta^*, z^*) - f(\theta^*, z^*)\{1 - \Pi 1(\theta^*, z^*)\} \right| \\
&\leq \left| \Pi f(\theta_n, z_n) - \Pi f(\theta^*, z^*) \right| + \left| f(\theta_n, z_n) - f(\theta^*, z^*) \right| + \left| \Pi 1(\theta_n, z_n) - \Pi 1(\theta^*, z^*) \right|
\end{aligned}$$

By continuity of $f$ we have $f(\theta_n, z_n) \to f(\theta^*, z^*)$ as $n \to \infty$. Since $1 \in C_b(\mathbb{R}^{d+1})$ it remains to show that $\Pi f$ is continuous for every $f \in C_b(\mathbb{R}^{d+1})$. Now

$$\begin{aligned}
&\left| \Pi f(\theta_n, z_n) - \Pi f(\theta^*, z^*) \right| \\
&= \left| \int f(\theta', z') \min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta_n; 0, \Sigma)} \frac{\nu(\theta_n - \theta')}{\nu(\theta' - \theta_n)} e^{z' - z_n}\right\} \nu(\theta' - \theta_n) g(\mathrm{d}z' \mid \bar\theta) \mathrm{d}\theta' \right. \qquad (2.17) \\
&\qquad \left. - \int f(\theta', z') \min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta^*; 0, \Sigma)} \frac{\nu(\theta^* - \theta')}{\nu(\theta' - \theta^*)} e^{z' - z^*}\right\} \nu(\theta' - \theta^*) g(\mathrm{d}z' \mid \bar\theta) \mathrm{d}\theta' \right| \\
&\leq \int \left| \nu(\theta' - \theta_n) - \nu(\theta' - \theta^*) \right| \mathrm{d}\theta' \qquad (2.18) \\
&\quad + \int \left| \min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta_n; 0, \Sigma)} \frac{\nu(\theta_n - \theta')}{\nu(\theta' - \theta_n)} e^{z' - z_n}\right\} - \min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta^*; 0, \Sigma)} \frac{\nu(\theta^* - \theta')}{\nu(\theta' - \theta^*)} e^{z' - z^*}\right\} \right| \nu(\theta' - \theta^*) g(\mathrm{d}z' \mid \bar\theta) \mathrm{d}\theta'. \\
&\qquad (2.19)
\end{aligned}$$

For (2.18), Assumption 2 implies $\nu(\theta' - \theta_n) \to \nu(\theta' - \theta^*)$ as $n \to \infty$ and hence Scheffé's lemma yields

$$\int \left| \nu(\theta' - \theta_n) - \nu(\theta' - \theta^*) \right| \mathrm{d}\theta' \to 0.$$

For (2.19), the map

$$(\theta, z) \mapsto \min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta; 0, \Sigma)} \frac{\nu(\theta - \theta')}{\nu(\theta' - \theta)} e^{z'-z}\right\}$$

is continuous for all $\theta', z'$ since it is just a composition of continuous functions. Hence,

$$\left|\min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta_n; 0, \Sigma)} \frac{\nu(\theta_n - \theta')}{\nu(\theta' - \theta_n)} e^{z'-z_n}\right\} - \min\left\{1, \frac{\varphi(\theta'; 0, \Sigma)}{\varphi(\theta^*; 0, \Sigma)} \frac{\nu(\theta^* - \theta')}{\nu(\theta' - \theta^*)} e^{z'-z^*}\right\}\right| \to 0$$

for every $(\theta', z')$ and an application of dominated convergence shows that (2.19) goes to zero. □

## S3. Proofs of Section 5

### S3·1. *Central Limit Theorem for Likelihood Estimators*

We detail here the proof of Theorem 3. For clarity we explicitly state the probability space supporting all random variables that are used to prove our limit theorem. For integers $N, T, k$ we introduce the space $E_T = \Theta \times \mathbb{R}^{NTk}$ where $\Theta \subset \mathbb{R}^d$ is the parameter space equipped with the Borel $\sigma$-algebra and probability measure $\mathbb{P}_T(\mathrm{d}\theta, \mathrm{d}u) = \pi_T^{\omega}(\mathrm{d}\theta) m_{T,\theta}(\mathrm{d}u)$. Finally, we will work with the Borel probability measure $\mathbb{P}$ on $E$ where $E = \mathsf{Y}^{\mathbb{N}} \times \prod_{T=1}^{\infty} E_T,\ \mathbb{P} = \mathbb{P}^Y \otimes \bigotimes_{T=1}^{\infty} \mathbb{P}_T$.

We are interested in the asymptotic distribution of the relative error of the log-likelihood

$$Z_T(\theta) = \log \widehat{p}(Y_{1:T} \mid \theta, U) - \log p(Y_{1:T} \mid \theta),$$

where $U \sim m_{T,\theta}(\cdot)$ or $U \sim \pi_T^{\omega}(\cdot \mid \theta)$. Indeed, we have $\mathcal{L}\mathrm{aw}\{Z_T(\theta)\} = g_T^{\omega}(\cdot \mid \theta)$ when $U \sim m_{T,\theta}(\cdot)$ and $\mathcal{L}\mathrm{aw}\{Z_T(\theta)\} = \bar{g}_T^{\omega}(\cdot \mid \theta)$ when $U \sim \pi_T^{\omega}(\cdot \mid \theta)$. Weak convergence results for $Z_T(\theta)$ have been established in Deligiannidis et al. (2018, Theorem 1) using a Taylor expansion. However, the CLTs introduced therein do not provide a bound on the Lipschitz metric $d_{\mathrm{BL}}$ and are not uniform in the parameter $\theta$ as required in Assumption 3. In order to obtain a uniform bound for all functions in $\mathrm{BL}(\mathbb{R})$ with $\|f\|_{\mathrm{BL}} \leq 1$ and all parameter values for some neighbourhood $B(\bar{\theta})$ we need to introduce further assumptions. We follow the approach in Deligiannidis et al. (2018) and write

$$\begin{aligned} Z_T(\theta) &= \sum_{t=1}^{T} \log\left\{1 + \frac{\widehat{p}(Y_t \mid \theta, U_t) - p(Y_t \mid \theta)}{p(Y_t \mid \theta)}\right\} \\ &= \sum_{t=1}^{T} \log\left\{1 + \frac{\epsilon_N(Y_t, \theta)}{\sqrt{N}}\right\} \end{aligned}$$

where

$$\epsilon_N(Y_t, \theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{\overline{w}(Y_t, U_{t,i}, \theta) - 1\right\},$$

$\overline{w}(Y_t, U_{t,i}, \theta)$ being a normalized importance weight defined in (10). Recall that

$$\sigma^2(y, \theta) = E\left\{\epsilon_T(y, \theta)^2\right\} = \mathrm{var}\left\{\overline{w}(y, U_{1,1}, \theta)\right\}, \quad \sigma^2(\theta) = E\left\{\sigma^2(Y_1, \theta)\right\}.$$

Here the number of particles, $N$, is scaled proportionally to the number of observations, that is $N = \lceil \gamma T \rceil$ for some $\gamma > 0$. In the following we will take $\gamma = 1$ (that is $N = T$) for simplicity and without loss of generality. In order to show convergence of the bounded Lipschitz metric uniformly in $\theta$, we will exploit the relation

$$\log(1 + x) = x - \frac{x^2}{2} + \int_0^x \frac{u^2}{1 + u} \mathrm{d}u,$$

where for $x < 0$ we use the convention

$$\int_0^x \frac{u^2}{1+u} \mathrm{d}u = -\int_x^0 \frac{u^2}{1+u} \mathrm{d}u.$$

We thus obtain

$$Z_T(\theta) = \frac{1}{\sqrt{T}} \sum_{t=1}^T \epsilon_T(Y_t, \theta) - \frac{1}{2T} \sum_{t=1}^T \epsilon_T(Y_t, \theta)^2 + \sum_{t=1}^T R_T(Y_t, \theta), \tag{3.1}$$

with

$$R_T(y, \theta) = \int_0^{\epsilon_T(y,\theta)/\sqrt{T}} \frac{u^2}{1+u} \mathrm{d}u. \tag{3.2}$$

We recall the following assumptions regarding the normalized weights.

*Assumption* 4. There exists a closed $\varepsilon$-ball $B(\bar{\theta})$ around $\bar{\theta}$ and a function $g$ such that the normalized weight $\overline{w}(y, U_{1,1}, \theta)$ defined in (10) satisfies for some $\Delta > 0$

$$\sup_{\theta \in B(\bar{\theta})} E\left\{\overline{w}(y, U_{1,1}, \theta)^{2+\Delta}\right\} \leq g(y),$$

where $U_{1,1} \sim h(\cdot \mid y, \theta)$ and $\mu(g) < \infty$. Additionally, $\theta \mapsto \sigma^2(y, \theta)$ is continuous in $\theta$ on $B(\bar{\theta})$ for all $y \in \mathsf{Y}$.

We can relate expectations of powers of $\epsilon_T(y, \theta)$ to that of $\overline{w}(y, U_{1,1}, \theta)$ in the following way.

LEMMA 7. *For any $k \geq 2$ and any $T \geq 1$*

$$E\left\{|\epsilon_T(y, \theta)|^k\right\} \leq c(k) \left[E\left\{\overline{w}(y, U_{1,1}, \theta))^k\right\} + 1\right]$$

*where $c(k)$ is a constant only depending on $k$.*

*Proof.* This is Lemma 2 in Deligiannidis et al. (2018). We repeat it here for convenience. It holds

$$\begin{aligned}
E\left\{|\epsilon_T(y, \theta)|^k\right\} &= E\left[\left|\frac{1}{\sqrt{T}} \sum_{i=1}^T \left\{\overline{w}(y, U_{1,i}, \theta) - 1\right\}\right|^k\right] \\
&\leq c_1(k) E\left[\left|\frac{1}{T} \sum_{i=1}^T \left\{\overline{w}(y, U_{1,i}, \theta) - 1\right\}^2\right|^{k/2}\right] \\
&\leq c_1(k) \frac{1}{T} \sum_{i=1}^T E\left\{\left|\overline{w}(y, U_{1,i}, \theta) - 1\right|^k\right\} \\
&\leq c_1(k) c_2(k) \left[E\left\{\overline{w}(y, U_{1,1}, \theta))^k\right\} + 1\right]
\end{aligned}$$

for some constants $c_1(k), c_2(k)$ by application of the Marcinkiewicz–Zygmund, Jensen and $c_r$-inequalities. □

As a result we have thus

$$\sup_{\theta \in B(\bar{\theta})} E\left\{|\epsilon_T(y, \theta)|^k\right\} \leq c(k) \sup_{\theta \in B(\bar{\theta})} \left[E\left\{\overline{w}(y, U_{1,1}, \theta))^k\right\} + 1\right] \tag{3.3}$$

and the left-hand-side is finite whenever the right-hand-side is finite.

S3·2. *Moment Conditions for Weak Convergence*

Denote $\mathcal{Y}_T$ the $\sigma-$algebra spanned by the data $Y_{1:T} = (Y_1, \ldots, Y_T)$ observed up to $T$.

THEOREM 3 (MOMENT CONDITIONS FOR UCLT). *Under Assumption* 4 *we have the following uniform central limit theorems*

*a)*

$$\sup_{\theta \in B(\bar\theta)} d_{\mathrm{BL}} \left[ g_T^{\omega}(\cdot \mid \theta), \varphi \left\{ \cdot; -\sigma^2(\theta)/2, \sigma^2(\theta) \right\} \mid \mathcal{Y}_T \right] \xrightarrow{\mathbb{P}} 0,$$

*and*

*b)*

$$\sup_{\theta \in B(\bar\theta)} d_{\mathrm{BL}} \left[ \bar g_T^{\omega}(\cdot \mid \theta), \varphi \left\{ \cdot; \sigma^2(\theta)/2, \sigma^2(\theta) \right\} \mid \mathcal{Y}_T \right] \xrightarrow{\mathbb{P}} 0.$$

We will need the following auxiliary results.

LEMMA 8. *Let* $S_T(\theta) = \sum_{i=1}^T \xi_i(\theta)$ *denote the sum of zero mean independent random variables* $\xi_1(\theta), \ldots, \xi_T(\theta)$ *such that* $\mathrm{var}(S_T) = 1$. *Then for any Lipschitz function* $f$ *with Lipschitz constant* $L$ *and* $Z \sim \mathcal{N}(0, 1)$

$$|E\left[f\left\{S_T(\theta)\right\} - f\left(Z\right)\right]| \leq L \left( 4E \left[ \sum_{i=1}^T \xi_i^2(\theta) 1_{\{|\xi_i(\theta)|>1\}} \right] + 3E \left[ \sum_{i=1}^T |\xi_i(\theta)|^3 1_{\{|\xi_i(\theta)|\leq 1\}} \right] \right).$$

*Proof.* This is Theorem 3.2 in Chen et al. (2010). □

The above result reduces the problem of showing weak convergence uniformly over some neighbourhood $B(\bar\theta)$ to uniform laws of large numbers for conditional higher order moments. Conditions to ensure uniformity in the convergence of averages are widely established. We will use the following result given in (Jennrich, 1969, Theorem 2).

LEMMA 9. *Let* $A \subset \mathbb{R}^d$ *be compact and let* $f : \mathbb{R}^k \times A \to \mathbb{R}$ *be continuous in* $\theta$ *for each* $y \in \mathbb{R}^k$ *and measurable in* $y$ *for each* $\theta \in A$. *Further assume that there exists an integrable function* $g$, *such that* $|f(y, \theta)| \leq g(y)$ *for all* $y$ *and* $\theta$. *For independent random variables* $Y_i \sim \mu$ $(i = 1, \ldots, T)$ *then* $\mathbb{P}^Y$-*almost surely*

$$\sup_{\theta \in A} \left| \frac{1}{T} \sum_{t=1}^T f\left(Y_t, \theta\right) - E\left\{f(Y_1, \theta)\right\} \right| \to 0,$$

*as* $T \to \infty$.

Before we proceed with the proof of Theorem 3, we note that Lemma 8 is not formulated in terms of conditional laws. However, considering conditionally (upon $\mathcal{Y}_T$) centred and independent random variables $\xi_{T,1}, \ldots, \xi_{T,T}$ such that $\sum_{i=1}^T \mathrm{var}\left\{\xi_i(\theta) | Y_{1:T}\right\} = 1$, we can apply the above lemma for every realization $Y_{1:T} = y_{1:T}$. Denote $P_T^y$ a regular conditional distribution associated with the law of $S_T = \xi_{T,1} + \ldots + \xi_{T,T}$ given $Y_{1:T} = y_{1:T}$. By applying Lemma 8, we get

$$\begin{aligned} & d_{\mathrm{BL}}\left\{P_T^y, \varphi(\,\cdot\,; 0, 1) \mid Y_{1:T} = y_{1:T}\right\} \\ & \leq 4E \left[ \sum_{i=1}^T \xi_i^2(\theta) 1_{\{|\xi_i(\theta)|>1\}} \mid Y_{1:T} = y_{1:T} \right] + 3E \left[ \sum_{i=1}^T |\xi_i(\theta)|^3 1_{\{|\xi_i(\theta)|\leq 1\}} \mid Y_{1:T} = y_{1:T} \right]. \end{aligned} \quad (3.4)$$

Thus, if the terms on the r.h.s. go to zero in $\mathbb{P}^Y$-probability then $d_{\mathrm{BL}}\left\{P_T^Y, \varphi(\,\cdot\,; 0, 1)\right\} \xrightarrow{\mathbb{P}^Y} 0$. With this reasoning we can apply Lemma 8 to prove Theorem 3.

*Proof of Theorem* 3, *part* a). Define

$$\xi_{T,t}(\theta) = \frac{\epsilon_T(Y_t, \theta)}{\sqrt{T}\sigma_T(Y_{1:T}, \theta)}, \quad S_T(\theta) = \sum_{t=1}^{T} \xi_{T,t}(\theta),$$

where

$$\sigma_T^2(Y_{1:T}, \theta) = \frac{1}{T} \sum_{t=1}^{T} \text{var}\left\{\epsilon_{T,t}(\theta) \mid \mathcal{Y}_T\right\}. \tag{3.5}$$

Thus

$$\text{var}\left\{S_T(\theta) \mid \mathcal{Y}_T\right\} = \sum_{t=1}^{T} \text{var}\left\{\xi_{T,t}(\theta)\right\} = 1.$$

In the following we will use the shorthand $\sigma_T(Y_{1:T}, \theta) = \left\{\sigma_T^2(Y_{1:T}, \theta)\right\}^{1/2}$ and $\sigma_T^r(Y_{1:T}, \theta) = \left\{\sigma_T^2(Y_{1:T}, \theta)\right\}^{r/2}$ for any real value $r$.

Then $S_T(\theta)$ fulfils the conditions of Lemma 8 conditionally on $\mathcal{Y}_T$. The random variable $Z_T(\theta)$ defined in (3.1) can be rewritten as

$$Z_T(\theta) = S_T(\theta)\sigma_T(Y_{1:T}, \theta) - \frac{1}{2T} \sum_{t=1}^{T} \epsilon_T(Y_t, \theta)^2 + \sum_{t=1}^{T} R_T(Y_t, \theta).$$

We have for $Z \sim \mathcal{N}(0, 1)$

$$\begin{aligned}
&\sup_{\theta \in B(\bar{\theta})} d_{\text{BL}}\left[\mathcal{L}aw\left\{Z_T(\theta)\right\}, \varphi\left\{\cdot; -\sigma^2(\theta)/2, \sigma^2(\theta)\right\} \mid \mathcal{Y}_T\right] \\
&= \sup_{\theta \in B(\bar{\theta})} d_{\text{BL}}\left[\mathcal{L}aw\left\{Z_T(\theta)\right\}, \mathcal{L}aw\left\{Z\sigma(\theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right] \\
&= \sup_{\theta \in B(\bar{\theta})} \sup_{\substack{f \in \text{BL}(\mathbb{R}) \\ \|f\|_{\text{BL}} \leq 1}} \left| E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T}, \theta) - \frac{1}{2T}\sum_{t=1}^{T}\epsilon_T(Y_t, \theta)^2 + \sum_{t=1}^{T} R_T(Y_t, \theta)\right\} \mid \mathcal{Y}_T\right] - E\left[f\left\{Z\sigma(\theta) - \frac{\sigma^2(\theta)}{2}\right\}\right]\right| \\
&\leq \sup_{\theta \in B(\bar{\theta})} \sup_{\substack{f \in \text{BL}(\mathbb{R}) \\ \|f\|_{\text{BL}} \leq 1}} \left| E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T}, \theta) - \frac{1}{2T}\sum_{t=1}^{T}\epsilon_T(Y_t, \theta)^2 + \sum_{t=1}^{T} R_T(Y_t, \theta) - \frac{\sigma^2(\theta)}{2} + \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right]\right. \qquad (3.6) \\
&\qquad \left. - E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T}, \theta) + \sum_{t=1}^{T} R_T(Y_t, \theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right]\right| \\
&\quad + \sup_{\theta \in B(\bar{\theta})} \sup_{\substack{f \in \text{BL}(\mathbb{R}) \\ \|f\|_{\text{BL}} \leq 1}} \left| E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T}, \theta) + \sum_{t=1}^{T} R_T(Y_t, \theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right]\right. \\
&\qquad \left. - E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T}, \theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right]\right| \qquad (3.7) \\
&\quad + \sup_{\theta \in B(\bar{\theta})} \sup_{\substack{f \in \text{BL}(\mathbb{R}) \\ \|f\|_{\text{BL}} \leq 1}} \left| E\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T}, \theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right] - E\left[f\left\{Z\sigma(\theta) - \frac{\sigma^2(\theta)}{2}\right\}\right]\right| \qquad (3.8)
\end{aligned}$$

Now we have for (3.6)

$$
\begin{aligned}
(3.6) \leq \sup_{\theta \in B(\bar{\theta})} \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}} \leq 1}} \Bigg| & E\left[ f\left\{ S_T(\theta)\sigma_T(Y_{1:T},\theta) - \frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2 + \sum_{t=1}^{T} R_T(Y_t,\theta) - \frac{\sigma^2(\theta)}{2} + \frac{\sigma^2(\theta)}{2} \right\} \mid \mathcal{Y}_T \right] \\
& - E\left[ f\left\{ S_T(\theta)\sigma_T(Y_{1:T},\theta) + \sum_{t=1}^{T} R_T(Y_t,\theta) - \frac{\sigma^2(\theta)}{2} \right\} \mid \mathcal{Y}_T \right] \Bigg| \\
\leq \sup_{\theta \in B(\bar{\theta})} & E\left[ \min\left\{ 1, \left| \frac{\sigma^2(\theta)}{2} - \frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2 \right| \right\} \mid \mathcal{Y}_T \right]
\end{aligned}
$$

where we use that $f$ is bounded and Lipschitz. We can bound this term by

$$
\begin{aligned}
& \sup_{\theta \in B(\bar{\theta})} E\left( \min\left\{ 1, \left| \frac{\sigma^2(\theta)}{2} - \frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2 \right| \right\} \Big| \mathcal{Y}_T \right) \\
& \leq \sup_{\theta \in B(\bar{\theta})} E\left( \min\left\{ 1, \left| \frac{1}{2T}\sum_{t=1}^{T} \left\{ \sigma^2(Y_t,\theta) - \sigma^2(\theta) \right\} \right| \right\} \Big| \mathcal{Y}_T \right) \qquad (3.9) \\
& \qquad + \sup_{\theta \in B(\bar{\theta})} E\left( \min\left\{ 1, \left| \frac{1}{2T}\sum_{t=1}^{T} \left\{ \epsilon_T(Y_t,\theta)^2 - \sigma^2(Y_t,\theta) \right\} \right| \right\} \Big| \mathcal{Y}_T \right).
\end{aligned}
$$

For any $0 < \delta < 1$, we can bound the first term on the r.h.s. of (3.9) by

$$
\begin{aligned}
& \sup_{\theta \in B(\bar{\theta})} E\left[ \min\left\{ 1, \left| \frac{1}{2T}\sum_{t=1}^{T} \left\{ \epsilon_T(Y_t,\theta)^2 - \sigma^2(Y_t,\theta) \right\} \right| \right\} \Big| \mathcal{Y}_T \right] \\
& \leq \sup_{\theta \in B(\bar{\theta})} \left( E\left[ \min\left\{ 1, \left| \frac{1}{2T}\sum_{t=1}^{T} \left\{ \epsilon_T(Y_t,\theta)^2 - \sigma^2(Y_t,\theta) \right\} \right|^{1+\delta} \right\} \Big| \mathcal{Y}_T \right] \right)^{\frac{1}{1+\delta}} \\
& \leq \left[ \frac{C}{2^{1+\delta}T^{1+\delta}} \sum_{t=1}^{T} \sup_{\theta \in B(\bar{\theta})} E\left\{ \left| \epsilon_T(Y_t,\theta)^2 - \sigma^2(Y_t,\theta) \right|^{1+\delta} \Big| \mathcal{Y}_T \right\} \right]^{\frac{1}{1+\delta}} \\
& \leq C\left[ \frac{C'}{2^{1+\delta}T^{1+\delta}} \sum_{t=1}^{T} \{1 + g(Y_t)\} \right]^{\frac{1}{1+\delta}} \to 0
\end{aligned}
$$

in $\mathbb{P}^Y$-probability by the law of large numbers using, in turn, Jensen's inequality, von Bahr–Esseen inequality (von Bahr & Esseen, 1965) as $E\left\{\epsilon_T(Y_t,\theta)^2 \mid \mathcal{Y}_T\right\} = \sigma^2(Y_t,\theta)$, $c_r$-inequality, (3.3) and Assumption 4 for $\Delta = 2\delta$, noting that

$$
\begin{aligned}
\sigma^2(Y_t,\theta) = E\left\{\epsilon_T(Y_t,\theta)^2 \mid \mathcal{Y}_T\right\} & \leq E\left\{ |\epsilon_T(Y_t,\theta)|^{2+\Delta} \right\}^{2/(2+\Delta)} \\
& \leq C \cdot \{g(Y_t) + 1\}^{2/(2+\Delta)},
\end{aligned}
$$

where the last inequality is due to (3.3). The second term on the right-hand side of (3.9) can be bounded

$$
\begin{aligned}
& \sup_{\theta \in B(\bar{\theta})} E\left( \min\left\{ 1, \left| \frac{1}{2T}\sum_{t=1}^{T} \left\{ \sigma^2(Y_t,\theta) - \sigma^2(\theta) \right\} \right| \right\} \mid \mathcal{Y}_T \right) \\
& \leq E\left( \min\left\{ 1, \sup_{\theta \in B(\bar{\theta})} \left| \frac{1}{2T}\sum_{t=1}^{T} \left\{ \sigma^2(Y_t,\theta) - \sigma^2(\theta) \right\} \right| \right\} \mid \mathcal{Y}_T \right).
\end{aligned}
$$

Noting that $\sigma^2(y,\theta)$ is continuous in $\theta$ for all $y$ by Assumption 4 and $\sigma^2(y,\theta) \le C \cdot \{1+g(y)\}^{2/(2+\Delta)}$ we can apply Lemma 9 to get

$$\sup_{\theta \in B(\bar\theta)} \left| \frac{1}{2T} \sum_{t=1}^{T} \left\{ \sigma^2(Y_t,\theta) - \sigma^2(\theta) \right\} \right| \xrightarrow{\mathbb{P}^Y} 0$$

and we can use dominated convergence to conclude that

$$E\left( E\left[ \min\left\{ 1, \sup_{\theta \in B(\bar\theta)} \left| \frac{1}{2T} \sum_{t=1}^{T} \left\{ \sigma^2(Y_t,\theta) - \sigma^2(\theta) \right\} \right| \right\} \mid \mathcal{Y}_T \right] \right) \to 0$$

and thus

$$E\left[ \min\left\{ 1, \sup_{\theta \in B(\bar\theta)} \left| \frac{1}{2T} \sum_{t=1}^{T} \left\{ \sigma^2(Y_t,\theta) - \sigma^2(\theta) \right\} \right| \right\} \mid \mathcal{Y}_T \right] \xrightarrow{\mathbb{P}^Y} 0.$$

The quantity (3.7) can be upper bounded by

$$(3.7) \le E\left[ \min\left\{ 1, \left| \sum_{t=1}^{T} R_T(Y_t,\theta) \right| \right\} \mid \mathcal{Y}_T \right] \le \sum_{t=1}^{T} E\left[ \min\left\{ 1, |R_T(Y_t,\theta)| \right\} \mid \mathcal{Y}_T \right]. \tag{3.10}$$

We will split the expectation into two terms

$$\begin{aligned} &E\left[ \min\{1, |R_T(Y_t,\theta)|\} \mid \mathcal{Y}_T \right] \\ &= E\left[ \min\{1, |R_T(Y_t,\theta)|\} 1_{\left\{ \left| \frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} \right| \le 1 \right\}} \mid \mathcal{Y}_T \right] + E\left[ \min\{1, |R_T(Y_t,\theta)|\} 1_{\left\{ \left| \frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} \right| > 1 \right\}} \mid \mathcal{Y}_T \right]. \end{aligned} \tag{3.11}$$

Recall

$$R_T(y,\theta) = \int_0^{\epsilon_T(y,\theta)/\sqrt{T}} \frac{u^2}{1+u} \mathrm{d}u.$$

We investigate the integral

$$\Psi(x) = \int_0^x \frac{u^2}{1+u} \mathrm{d}u \tag{3.12}$$

in more detail (see also Figure 1), where in the case $x < 0$, we interpret the above as an integral over the interval $[x, 0]$. Without loss of generality, we can always select $0 < \Delta < 1$ in Assumption 4. On the interval $(-1, 1]$, we can bound the function

$$\frac{u^2}{1+u} \le \frac{|u|^{1+\Delta}}{1+u},$$

as $0 < \Delta < 1$ where we show $\Delta = 0.1$ as an example in Figure 1. Subsequently, we bound for $x \in (-1, 1]$

$$\left| \int_0^x \frac{u^2}{1+u} \mathrm{d}u \right| \le \left| \int_0^x \frac{|u|^{1+\Delta}}{1+u} \mathrm{d}u \right| \le \left| x \cdot \frac{|x|^{1+\Delta}}{1+x} \right|,$$

i.e. the box containing the area under the curve. This is visualized in Figure 1. The integral (shaded blue) is bounded by the striped box. Hence, on the set $|\epsilon_T(y,\theta)/\sqrt{T}| \le 1$, we have

$$\left| \int_0^{\epsilon_T(y,\theta)/\sqrt{T}} \frac{u^2}{1+u} \mathrm{d}u \right| \le \left| \frac{|\epsilon_T(y,\theta)|^{2+\Delta}}{T^{1+\Delta/2}} \frac{1}{1+\epsilon_T(y,\theta)/\sqrt{T}} \right|.$$

For any non-negative random variable $X$ and event $A$, we have the identity

$$E\{\min(1, X) 1_A\} \le E\left\{ X 1_{X \le 1} 1_A \right\} + \mathbb{P}(X > 1),$$

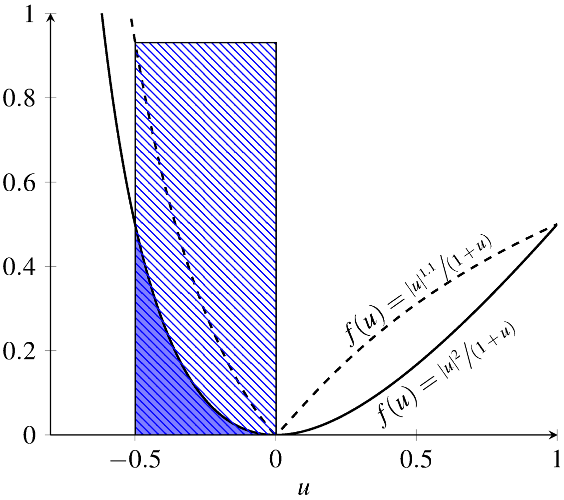


Fig. 1. For $x \in (-1, 1]$, $x = -0.5$ on the graph, the remainder of our expansion is estimated by the integral under the solid curve (blue shaded area). We bound this integral first by dashed line and then we approximate the integral by the box containing the area (lines).

so we can bound the first term on the right-hand side of (3.11) for every $t = 1, \dots, T$

$$
\begin{aligned}
&E\left[\min\{1, |R_T(Y_t,\theta)|\}\, 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}\right|\le 1\right\}} \mid \mathcal{Y}_T\right] \\
&\le E\left[\left|\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\right| 1_{\left\{\left|\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\right|\le 1\right\}} 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}\right|\le 1\right\}} \middle| \mathcal{Y}_T\right] \\
&\quad + \mathbb{P}\left\{\left|\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\right| > 1 \middle| \mathcal{Y}_T\right\}.
\end{aligned}
$$ 510

By inspection of the function, similarly to before,

$$u \mapsto \frac{|u|^{2+\Delta}}{1+u}$$

one can easily verify that there exist $0 < \delta_1 < 1$ and $\delta_2 > 0$ such that

$$\frac{|u|^{2+\Delta}}{1+u} \le 1 \Leftrightarrow -\delta_1 \le u \le \delta_2.$$

Thus we have

$$
\begin{aligned}
&E\left[\left|\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\right| 1_{\left\{\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\le 1\right\}} 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}\right|\le 1\right\}} \middle| \mathcal{Y}_T\right] \\
&\le E\left[\left|\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\right| 1_{\{-\delta_1\le\epsilon_T(Y_t,\theta)/\sqrt{T}\le\delta_2\}} \middle| \mathcal{Y}_T\right] \\
&\le \frac{1}{(1-\delta_1)T^{1+\Delta/2}} E\left[|\epsilon_T(Y_t,\theta)|^{2+\Delta} \middle| \mathcal{Y}_T\right], \qquad (3.13)
\end{aligned}
$$

515

while

$$
\begin{aligned}
&\mathbb{P}\left\{\left|\frac{\epsilon_T^{2+\Delta}(Y_t,\theta)/T^{1+\Delta/2}}{1+\epsilon_T(Y_t,\theta)/\sqrt{T}}\right| > 1\middle|\mathcal{Y}_T\right\} \\
&\leq \mathbb{P}\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{T^{1/2}}\right| > \min\{\delta_1,\delta_2\}\middle|\mathcal{Y}_T\right\} \\
&\leq \frac{1}{\min\{\delta_1,\delta_2\}^{2+\Delta}T^{1+\Delta/2}} E\left[|\epsilon_T(Y_t,\theta)|^{2+\Delta}\middle|\mathcal{Y}_T\right]. \qquad (3.14)
\end{aligned}
$$

The second term on the right-hand side of (3.11) is bounded by

$$
\begin{aligned}
&E\left[\min\{1,|R_T(Y_t,\theta)|\}\,1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}\right|>1\right\}} \mid \mathcal{Y}_T\right] \\
&\leq E\left[|R_T(Y_t,\theta)|\,1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}\right|>1\right\}} \mid \mathcal{Y}_T\right]. \qquad (3.15)
\end{aligned}
$$

As $\epsilon_T(Y_t,\theta)/\sqrt{T} \geq -1$, (3.15) is null for $\epsilon_T(Y_t,\theta) < -1$ so writing $X^+ = \max\{0, X\}$ this can be rewritten as

$$
\begin{aligned}
&E\left\{\int_0^{\epsilon_T(Y_t,\theta)/\sqrt{T}} \frac{u^2}{1+u}\mathrm{d}u\,1_{\{\epsilon_T(Y_t,\theta)/\sqrt{T}\geq 1\}}\middle|\mathcal{Y}_T\right\} \\
&\leq E\left[\int_0^{(\epsilon_T(Y_t,\theta)/\sqrt{T})^+} \frac{u^2}{1+u}\mathrm{d}u\middle|\mathcal{Y}_T\right] \\
&= \int_0^\infty \frac{u^2}{1+u}\mathbb{P}\left\{\epsilon_T(Y_t,\theta)^+ > \sqrt{T}u\middle|\mathcal{Y}_T\right\}\mathrm{d}u,
\end{aligned}
$$

where we have used that for the function (3.12) is increasing and differentiable on its domain so

$$
E\{\Psi(|X|)\} = \Psi(0) + \int_0^\infty \Psi'(u)P(|X| > u)\mathrm{d}u.
$$

For $\Delta \in (0,1)$, we bound the remainder using

$$
\begin{aligned}
&= \int_0^\infty \frac{u^2}{1+u}\mathbb{P}\left\{\epsilon_T(Y_t,\theta)^+ > \sqrt{T}u\middle|\mathcal{Y}_T\right\}\mathrm{d}u \\
&\leq \int_0^\infty \frac{u^2}{1+u}\frac{E\left\{|\varepsilon_T(Y_t,\theta)|^{2+\Delta}\middle|\mathcal{Y}_T\right\}}{T^{(2+\Delta)/2}u^{2+\Delta}}\mathrm{d}u \\
&= \int_0^\infty \frac{1}{(1+u)u^\Delta}\mathrm{d}u\frac{1}{T^{1+\Delta/2}}E\left\{|\varepsilon_T(Y_t,\theta)|^{2+\Delta}\middle|\mathcal{Y}_T\right\} \\
&= C(\Delta)\frac{E\left\{|\varepsilon_T(Y_t,\theta)|^{2+\Delta}\middle|\mathcal{Y}_T\right\}}{T^{1+\Delta/2}} \qquad (3.16)
\end{aligned}
$$

noting that

$$
\int_0^\infty \frac{1}{(1+u)u^\Delta}du = C(\Delta) < \infty
$$

for $\Delta \in (0,1)$. Hence we can bound (3.11) by the sum of (3.13), (3.14) and (3.16) so, by using (3.10), we obtain a bound for (3.7)

$$
(3.7) \leq \frac{1}{(1-\delta_1)T^{(1+\Delta)/2}}\sum_{t=1}^T \sup_{\theta\in B(\bar{\theta})} E\left\{|\varepsilon_T(Y_t,\theta)|^{2+\Delta}\middle|\mathcal{Y}_T\right\}
$$

$$+ \frac{1}{\min\{\delta_1, \delta_2\}^{2+\Delta} T^{(1+\Delta)/2}} \sum_{t=1}^{T} \sup_{\theta \in B(\bar{\theta})} E\left\{ |\varepsilon_T(Y_t, \theta)|^{2+\Delta} \,\Big|\, \mathcal{Y}_T \right\}$$

$$+ C(\Delta) \frac{1}{T^{1+\Delta/2}} \sum_{t=1}^{T} \sup_{\theta \in B(\bar{\theta})} E\left\{ |\varepsilon_T(Y_t, \theta)|^{2+\Delta} \,\Big|\, \mathcal{Y}_T \right\} \to 0$$

which all converge in $\mathbb{P}^Y$-probability by (3.3), Assumption 4 and the law of large numbers.

We are now going to bound (3.8). We will use the fact that any constant $c$ and any two random variables $X_1, X_2$ we have for $c > 0$

$$\begin{aligned}
\sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}} \le 1}} |E[f(cX_1) - f(cX_2)]| &\le \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}} \le c}} |E[f(X_1) - f(X_2)]| \\
&= \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}} \le c}} \left| E\left[ c \left\{ \frac{f(X_1)}{c} - \frac{f(X_2)}{c} \right\} \right] \right| \\
&\le c \cdot \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{L}} \le 1}} |E[\{f(X_1) - f(X_2)\}]| .
\end{aligned}$$

Note that we only require $\|f\|_{\mathrm{L}} \le 1$ ($\|f\|_{\mathrm{L}}$ denoting the Lipschitz constant) in the last line alleviating the bound on the supremum $\|f\|_{\infty}$. The aim of the following paragraphs is to apply the above inequality and Lemma 8 to find a bound on (3.8). Omitting for the moment the supremum over the set $B(\bar{\theta})$ we compute for (3.8)

$$\begin{aligned}
&\sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}} \le 1}} |E[f\{S_T(\theta)\sigma_T(Y_{1:T}, \theta)\} \mid \mathcal{Y}_T] - E[f\{Z\sigma(\theta)\} \mid \mathcal{Y}_T]| \\
&\le \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}} \le 1}} |E[f\{S_T(\theta)\sigma_T(Y_{1:T}, \theta)\} \mid \mathcal{Y}_T] - E[f\{Z\sigma_T(Y_{1:T}, \theta)\} \mid \mathcal{Y}_T]| \\
&\quad + \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{BL}} \le 1}} |E[f\{Z\sigma_T(Y_{1:T}, \theta)\} \mid \mathcal{Y}_T] - E[f\{Z\sigma(\theta)\}]| \\
&\le \sigma_T(Y_{1:T}, \theta) \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{L}} \le 1}} |E[f\{S_T(\theta)\} \mid \mathcal{Y}_T] - E[f\{Z\}]| + E[|Z|]\, |\sigma_T(Y_{1:T}, \theta) - \sigma(\theta)| \\
&\le \sigma_T(Y_{1:T}, \theta) \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{L}} \le 1}} |E[f\{S_T(\theta)\} \mid \mathcal{Y}_T] - E[f\{Z\}]| + \left(\frac{2}{\pi}\right)^{1/2} |\sigma_T(Y_{1:T}, \theta) - \sigma(\theta)| , \qquad (3.17)
\end{aligned}$$

We have already shown

$$\sup_{\theta \in B(\bar{\theta})} \left| \sigma_T^2(Y_{1:T}, \theta) - \sigma^2(\theta) \right| = \sup_{\theta \in B(\bar{\theta})} \left| \sum_{t=1}^{T} \frac{\sigma^2(Y_t, \theta)}{T} - \sigma^2(\theta) \right| \xrightarrow{\mathbb{P}^Y} 0,$$

by the uniform law of large numbers (Lemma 9). Using $|\sqrt{a} - \sqrt{b}| \le \sqrt{|a-b|}$, we have

$$\sup_{\theta \in B(\bar{\theta})} |\sigma_T(Y_{1:T}, \theta) - \sigma(\theta)| \xrightarrow{\mathbb{P}^Y} 0.$$

For the first part of (3.17), by Lemma 8 applied conditionally on $\mathcal{Y}_T$

$$\sup_{\theta \in B(\bar{\theta})} \sigma_T(Y_{1:T}, \theta) \sup_{\substack{f \in \mathrm{BL}(\mathbb{R}) \\ \|f\|_{\mathrm{L}} \le 1}} \left| E\left[ f\left\{ \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\epsilon_T(Y_t, \theta)}{\sigma_T(Y_{1:T}, \theta)} \right\} \mid \mathcal{Y}_T \right] - E[f\{Z\}] \right|$$

$$\leq 4 \sup_{\theta \in B(\bar{\theta})} \sigma_T(Y_{1:T}, \theta) \sum_{t=1}^{T} E\left[\left\{\frac{\epsilon(Y_t, \theta)}{\sqrt{T}\sigma_T(Y_{1:T}, \theta)}\right\}^2 1_{\left\{\frac{|\epsilon_T(Y_t,\theta)|}{\sqrt{T}\sigma_T(Y_{1:T},\theta)} > 1\right\}} \mid \mathcal{Y}_T\right] \tag{3.18}$$

$$+ 3 \sup_{\theta \in B(\bar{\theta})} \sigma_T(Y_{1:T}, \theta) \sum_{i=1}^{T} E\left[\left|\frac{\epsilon(Y_t, \theta)}{\sqrt{T}\sigma_T(Y_{1:T}, \theta)}\right|^3 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}\right| \leq 1\right\}} \mid \mathcal{Y}_T\right]. \tag{3.19}$$

In order to control the $\sigma^2(Y_{1:T}, \theta)$ term consider the set

$$A_T(\delta) = \left\{y_{1:T} : \sup_{\theta \in B(\bar{\theta})} \left|\sigma_T^2(y_{1:T}, \theta) - \sigma^2(\theta)\right| \leq \delta\right\}.$$

The uniform convergence of $\sigma_T^2(Y_{1:T}, \theta)$ means that for any $\delta > 0$

$$\mathbb{P}^Y \left\{A_T(\delta)^{\complement}\right\} \to 0$$

as $T \to \infty$. Choosing $\delta > 0$ for any family of random variables $\gamma_T(Y_{1:T}, \theta)$ we have

$$\mathbb{P}^Y\left(\left|\sup_{\theta \in B(\bar{\theta})} \gamma_T(Y_{1:T}, \theta)\right| > \delta\right)$$
$$= \mathbb{P}^Y\left(\left\{\left|\sup_{\theta \in B(\bar{\theta})} \gamma_T(Y_{1:T}, \theta)\right| > \delta\right\} \cap A_T(\delta)\right) + \mathbb{P}^Y\left(\left\{\left|\sup_{\theta \in B(\bar{\theta})} \gamma_T(Y_{1:T}, \theta)\right| > \delta\right\} \cap A_T(\delta)^{\complement}\right)$$

where we have already shown

$$\mathbb{P}^Y\left[\left\{\left|\sup_{\theta \in B(\bar{\theta})} \gamma_T(Y_{1:T}, \theta)\right| > \eta\right\} \cap A_T(\delta)^{\complement}\right] \leq \mathbb{P}^Y\left\{A_T(\delta)^{\complement}\right\} \to 0. \tag{3.20}$$

Hence, for showing the convergence in probability for a random variable $\gamma_T(Y_{1:T}, \theta)$ it suffices to ensure convergence on the set $A(\delta)$. On the set $A(\delta)$ we can estimate $\sigma_T^2\{Y_{1:T}(\omega), \theta\} \geq \sigma^2(\theta) - \delta$ for all $\theta$. By continuity of $\sigma^2(\theta)$—and by shrinking $B(\bar{\theta})$ if necessary—we further have $\sigma^2(\theta) \geq \sigma^2(\bar{\theta}) - \delta$ for all $\theta \in B(\bar{\theta})$ and we get for (3.18), ignoring the constant for now

$$\sup_{\theta \in B(\bar{\theta})} \sigma_T(Y_{1:T}, \theta) \sum_{t=1}^{T} E\left[\left\{\frac{\epsilon(Y_t, \theta)}{\sqrt{T}\sigma_T(Y_{1:T}, \theta)}\right\}^2 1_{\left\{\frac{|\epsilon_T(Y_t,\theta)|}{\sqrt{T}\sigma_T(Y_{1:T},\theta)} > 1\right\}} \mid \mathcal{Y}_T\right] \mathbf{1}_{A_T(\delta)}$$
$$\leq \sup_{\theta \in B(\bar{\theta})} \frac{1}{\sigma_T^{1+\Delta}(Y_{1:T}, \theta)} \sum_{t=1}^{T} E\left[\left\{\frac{\epsilon(Y_t, \theta)}{\sqrt{T}}\right\}^{2+\Delta} 1_{\left\{\frac{|\epsilon_T(Y_t,\theta)|}{\sqrt{T}\sigma_T(Y_{1:T},\theta)} > 1\right\}} \mid \mathcal{Y}_T\right] \mathbf{1}_{A_T(\delta)}$$
$$\leq \sup_{\theta \in B(\bar{\theta})} \frac{1}{\left\{\sigma^2(\theta) - \delta\right\}^{(1+\Delta)/2}} \frac{1}{T^{1+\Delta/2}} \sum_{t=1}^{T} E\left\{|\epsilon_T(Y_t, \theta)|^{2+\Delta} \mid \mathcal{Y}_T\right\} \mathbf{1}_{A_T(\delta)}$$
$$\leq \frac{C}{\left\{\sigma^2(\bar{\theta}) - 2\delta\right\}^{(1+\Delta)/2} T^{1+\Delta/2}} \sum_{t=1}^{T} \{g(Y_t) + 1\} \xrightarrow{\mathbb{P}^Y} 0$$

independently of $\theta$ by the Marcinkiewicz-Zygmund law of large numbers (Kallenberg, 2006, Theorem 4.23). Together with (3.20) we can conclude that (3.18), vanishes in probability.

The second part, (3.19), can be controlled similarly via

$$\sigma_T(Y_{1:T}, \theta) \sum_{i=1}^{T} E\left[\left|\frac{\epsilon(Y_t, \theta)}{\sqrt{T}\sigma_T(Y_{1:T}, \theta)}\right|^3 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}\right| \leq 1\right\}} \mid \mathcal{Y}_T\right] \mathbf{1}_{A_T(\delta)}$$

$$
\begin{aligned}
&\leq \frac{1}{\sigma_T^{1+\Delta}(Y_{1:T},\theta)} \sum_{t=1}^{T} E\left[\left|\frac{\epsilon(Y_t,\theta)}{\sqrt{T}}\right|^{2+\Delta} 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}\right|\leq 1\right\}} \mid \mathcal{Y}_T\right] \mathbf{1}_{A_T(\delta)} \\
&\leq \frac{1}{\left\{\sigma^2(\theta)-\delta\right\}^{(1+\Delta)/2}} \sum_{t=1}^{T} E\left[\left|\frac{\epsilon(Y_t,\theta)}{\sqrt{T}}\right|^{2+\Delta} 1_{\left\{\left|\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}\sigma_T(Y_{1:T},\theta)}\right|\leq 1\right\}} \mid \mathcal{Y}_T\right] \mathbf{1}_{A_T(\delta)} \\
&\leq \frac{1}{\left\{\sigma^2(\bar{\theta})-2\delta\right\}^{(1+\Delta)/2}} \frac{1}{T^{1+\Delta/2}} \sum_{t=1}^{T} E\left\{|\epsilon(Y_t,\theta)|^{2+\Delta} \mid \mathcal{Y}_T\right\} \mathbf{1}_{A_T(\delta)} \\
&\leq \frac{C}{\left\{\sigma^2(\bar{\theta})-2\delta\right\}^{(1+\Delta)/2} T^{1+\Delta/2}} \sum_{t=1}^{T} \{g(Y_t)+1\} \xrightarrow{\mathbb{P}^Y} 0,
\end{aligned}
$$

which also does not depend on $\theta$. A similar argument to the one used to conclude in the case of (3.18) suffices also in this case. □

Turning to part *b)*, we analyse $Z_T(\theta)$ under stationarity. Therefore we need to introduce the probability measure of the auxiliary variables under stationarity, i.e. the distribution of the auxiliary variables conditional on the current state $\theta$. The conditional density is given by

$$
\pi(u \mid \theta) = \frac{\pi(u,\theta)}{\pi(\theta)} = \pi(\theta)\frac{\hat{p}(y \mid \theta, u)}{p(y \mid \theta)} m(u)/\pi(\theta) = \frac{\hat{p}(y \mid \theta, u)}{p(y \mid \theta)} m(u)
$$

which gives us the Radon-Nikodym derivative

$$
\frac{d\pi(\cdot \mid \theta)}{dm} = \prod_{t=1}^{T} \frac{\hat{p}(y_t \mid \theta, u_t)}{p(y_t \mid \theta)} = \exp\{Z_T(\theta)\}
$$

or alternatively

$$
\begin{aligned}
\prod_{t=1}^{T} \frac{\hat{p}(y_t \mid \theta, u_t)}{p(y_t \mid \theta)} &= \prod_{t=1}^{T} \left\{\frac{\hat{p}(y_t \mid \theta, u_t) - p(y_t \mid \theta)}{p(y_t \mid \theta)} + 1\right\} \\
&= \prod_{t=1}^{T} \left\{\frac{\epsilon_T(y_t,\theta)}{\sqrt{T}} + 1\right\}.
\end{aligned}
$$

The limiting distribution will now be Gaussian with a shifted mean, i.e. $\varphi(\cdot;\sigma^2(\theta)/2,\sigma^2(\theta))$. For $Z \sim \mathcal{N}(0,1)$ we will make use of the following identity

$$
E\left\{f\left(Z\sigma + \frac{\sigma^2}{2}\right)\right\} = E\left\{f\left(Z\sigma - \frac{\sigma^2}{2}\right)\exp\left(Z\sigma - \frac{\sigma^2}{2}\right)\right\}
$$

for every bounded Lipschitz function $f$. The identity is not restricted to this case, but we will only consider bounded Lipschitz functions. Before we present the proof, we have the following useful result.

PROPOSITION 4. *The Radon-Nikodym derivative is asymptotically uniformly bounded in its second moment,*

$$
\limsup_{T\to\infty} \sup_{\theta\in B(\bar{\theta})} E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1\right\}^2 \mid \mathcal{Y}_T\right] < \infty.
$$

*Proof.* Using independence of $\left(U_{t,1:T}\right)_{t\geq 1}$ we compute for all $\theta \in B(\bar{\theta})$

$$
\prod_{t=1}^{T} E\left\{\frac{\epsilon_T(Y_t,\theta)^2}{T} + 2\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1 \mid \mathcal{Y}_T\right\} = \prod_{t=1}^{T}\left\{\frac{\sigma^2(Y_t,\theta)}{T} + 1\right\}
$$

$$
\begin{aligned}
&\leq \exp\left\{\sum_{t=1}^{T} \frac{\sigma^2(Y_t,\theta)}{T}\right\}\\
&\leq \exp\left\{\sum_{t=1}^{T} \frac{C(1+g(Y_t))^{2/(2+\Delta)}}{T}\right\}\\
&\to \exp\left[CE\left\{(1+g\,(Y_1))^{2/(2+\Delta)}\right\}\right]
\end{aligned}
$$

in $\mathbb{P}^Y$-probability, which is clearly finite by Assumption 4. □

In the following we denote $E$ the expectation under $m$ and $\tilde{E}$ the expectation under $\pi(\cdot \mid \theta)$. Using the Radon-Nikodym derivative, it is possible to relate the expectation of $\epsilon_T(y,\theta)^k$ under $U$ at stationarity (conditional on $\theta$) to the expectation under $U \sim m(\cdot)$ by

$$
E_{U\sim\pi(\cdot|\theta)}\left\{\epsilon_T(y,\theta)^k\right\} = \frac{1}{\sqrt{T}} E_{U\sim m(\cdot)}\left\{\epsilon_T(y,\theta)^{k+1}\right\} + E_{U\sim m(\cdot)}\left\{\epsilon_T(y,\theta)^k\right\};
$$

see (Deligiannidis et al., 2018, Lemma 4) for a proof. We are now able to prove the second part of Theorem 3.

*Proof of Theorem* 3*, part* b). Again we take $Z \sim \mathcal{N}(0,1)$ and use the same decomposition as before, but with all expectations replaced by $\tilde{E}$, the expectation at stationarity:

$$
\begin{aligned}
&\sup_{\theta\in B(\bar\theta)} d_{\mathrm{BL}}\left[\bar g_T^{\omega}(\cdot \mid \theta), \varphi\left\{\cdot;\sigma^2(\theta)/2, \sigma^2(\theta)\right\}\right]\\
&= \sup_{\theta\in B(\bar\theta)} \sup_{\substack{f\in \mathrm{BL}(\mathbb{R})\\ \|f\|_{\mathrm{BL}}\leq 1}} \Bigg|\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta) - \frac{1}{2T}\sum_{t=1}^{T}\epsilon_T(Y_t,\theta)^2 + \sum_{t=1}^{T} R_T(Y_t,\theta)\right\} \mid \mathcal{Y}_T\right]\\
&\qquad - \tilde{E}\left[f\left\{Z\sigma(\theta) + \frac{\sigma^2(\theta)}{2}\right\}\right]\Bigg|\\
&\leq \sup_{\theta\in B(\bar\theta)} \sup_{\substack{f\in \mathrm{BL}(\mathbb{R})\\ \|f\|_{\mathrm{BL}}\leq 1}} \Bigg|\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta) - \frac{1}{2T}\sum_{t=1}^{T}\epsilon_T(Y_t,\theta)^2 + \sum_{t=1}^{T} R_T(Y_t,\theta) - \frac{\sigma^2(\theta)}{2} + \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right] \qquad (3.21)\\
&\qquad -\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta) + \sum_{t=1}^{T} R_T(Y_t,\theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right]\Bigg|\\
&\quad + \sup_{\theta\in B(\bar\theta)} \sup_{\substack{f\in \mathrm{BL}(\mathbb{R})\\ \|f\|_{\mathrm{BL}}\leq 1}} \Bigg|\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta) + \sum_{t=1}^{T} R_T(Y_t,\theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right] \qquad (3.22)\\
&\qquad - \tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right]\Bigg|\\
&\quad + \sup_{\theta\in B(\bar\theta)} \sup_{\substack{f\in \mathrm{BL}(\mathbb{R})\\ \|f\|_{\mathrm{BL}}\leq 1}} \Bigg|\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta) - \frac{\sigma^2(\theta)}{2}\right\} \mid \mathcal{Y}_T\right] - \tilde{E}\left[f\left\{Z\sigma(\theta) + \frac{\sigma^2(\theta)}{2}\right\}\right]\Bigg|. \qquad (3.23)
\end{aligned}
$$

For (3.21) we have

$$(3.21) \leq \sup_{\theta \in B(\bar{\theta})} \tilde{E}\left[\min\left\{1, \left|\frac{\sigma^2(\theta)}{2} - \frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2\right|\right\} \mid \mathcal{Y}_T\right].$$

An application of Cauchy-Schwartz yields

$$\begin{aligned}
&\tilde{E}\left[\min\left\{1, \left|\frac{\sigma^2(\theta)}{2} - \frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2\right|\right\} \mid \mathcal{Y}_T\right]\\
&= E\left[\min\left\{1, \left|\frac{\sigma^2(\theta)}{2} - \frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2\right|\right\} \prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1\right\} \mid \mathcal{Y}_T\right]\\
&\leq E\left[\min\left\{1, \left|\frac{\sigma^2(\theta)}{2} - \frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2\right|^2\right\} \mid \mathcal{Y}_T\right]^{1/2} E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1\right\}^2 \mid \mathcal{Y}_T\right]^{1/2}\\
&\leq E\left[\min\left\{1, \left|\frac{\sigma^2(\theta)}{2} - \frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2\right|\right\} \mid \mathcal{Y}_T\right]^{1/2} E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1\right\}^2 \mid \mathcal{Y}_T\right]^{1/2}.
\end{aligned}$$

By Proposition 4

$$\limsup_{T\to\infty} \sup_{\theta \in B(\bar{\theta})} E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1\right\}^2 \mid \mathcal{Y}_T\right]^{1/2} < \infty$$

and we have previously shown that

$$\sup_{\theta \in B(\bar{\theta})} E\left[\min\left\{1, \left|-\frac{1}{2T}\sum_{t=1}^{T} \epsilon_T(Y_t,\theta)^2 + \frac{\sigma^2(\theta)}{2}\right|\right\} \mid \mathcal{Y}_T\right] \to 0.$$

As for the remainder (3.22) we argue analogously

$$\begin{aligned}
&\tilde{E}\left[\min\left\{1, \left|\sum_{t=1}^{T} R_T(Y_t,\theta)\right|\right\} \mid \mathcal{Y}_T\right]\\
&= E\left[\min\left\{1, \left|\sum_{t=1}^{T} R_T(Y_t,\theta)\right|\right\} \prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1\right\} \mid \mathcal{Y}_T\right]\\
&\leq E\left[\min\left\{1, \left|\sum_{t=1}^{T} R_T(Y_t,\theta)\right|\right\}^2 \mid \mathcal{Y}_T\right]^{1/2} \cdot E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1\right\}^2 \mid \mathcal{Y}_T\right]^{1/2}\\
&\leq E\left[\min\left\{1, \left|\sum_{t=1}^{T} R_T(Y_t,\theta)\right|\right\} \mid \mathcal{Y}_T\right]^{1/2} \cdot E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1\right\}^2 \mid \mathcal{Y}_T\right]^{1/2}.
\end{aligned}$$

The first factor vanishes in probability as we have shown in the proof of Theorem 3(a), where as the second factor is bounded by Proposition 4.

For (3.23), note first that

$$E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}} + 1\right\} \mid \mathcal{Y}_T\right] = 1 \quad \text{and} \quad E\left[e^{Z\sigma(\theta) - \frac{\sigma^2(\theta)}{2}}\right] = 1.$$

Hence, we can write

$$
\begin{aligned}
&\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}\mid \mathcal{Y}_T\right]\\
&=\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}\mid \mathcal{Y}_T\right]E\left[e^{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}}\right]\\
&=\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}e^{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}}\mid \mathcal{Y}_T\right]
\end{aligned}
$$

and similarly

$$
\begin{aligned}
\tilde{E}\left[f\left\{Z\sigma(\theta)+\frac{\sigma^2(\theta)}{2}\right\}\right]&=E\left[f\left\{Z\sigma(\theta)+\frac{\sigma^2(\theta)}{2}\right\}e^{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}}\right]E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}\mid \mathcal{Y}_T\right]\\
&=E\left[f\left\{Z\sigma(\theta)+\frac{\sigma^2(\theta)}{2}\right\}e^{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}}\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}\mid \mathcal{Y}_T\right],
\end{aligned}
$$

where we used that $Z$ is independent of all other random variables in both cases. Using these identities we obtain

$$
\begin{aligned}
&\left|\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}\mid \mathcal{Y}_T\right]-\tilde{E}\left[f\left\{Z\sigma(\theta)+\frac{\sigma^2(\theta)}{2}\right\}\right]\right|\\
&\leq\left|\tilde{E}\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}-f\left\{Z\sigma(\theta)+\frac{\sigma^2(\theta)}{2}\right\}\mid \mathcal{Y}_T\right]\right|\\
&\leq\left|E\left[\left(f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}-f\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}\right)\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}e^{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}}\mid \mathcal{Y}_T\right]\right|\\
&\leq\left|E\left(\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}-f\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}\right]^2\mid \mathcal{Y}_T\right)^{\frac{1}{2}}\right|\\
&\times\left|E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}^2e^{2\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}}\mid \mathcal{Y}_T\right]^{\frac{1}{2}}\right|.
\end{aligned}
$$

We investigate the two factors of the product separately. First we use the fact that $\|f\|_\infty\leq 1$ when $\|f\|_{\mathrm{BL}}\leq 1$ (see (1.3)) and thus

$$
\begin{aligned}
&\left|\sup_{\theta\in B(\bar{\theta})}\sup_{\substack{f\in\mathrm{BL}(\mathbb{R})\\ \|f\|_{\mathrm{BL}}\leq 1}}E\left(\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}-f\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}\right]^2\mid \mathcal{Y}_T\right)^{\frac{1}{2}}\right|\\
&\leq\sup_{\theta\in B(\bar{\theta})}\sup_{\substack{f\in\mathrm{BL}(\mathbb{R})\\ \|f\|_{\mathrm{BL}}\leq 1}}\left|E\left(\left[f\left\{S_T(\theta)\sigma_T(Y_{1:T},\theta)-\frac{\sigma^2(\theta)}{2}\right\}-f\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}\right]\mid \mathcal{Y}_T\right)^{\frac{1}{2}}\right|\to 0
\end{aligned}
$$

in $\mathbb{P}^Y$-probability as established in the previous part. For the second factor note that $Z$ is independent of all other random variables and hence

$$\sup_{\theta\in B(\bar\theta)}\left| E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}^2 e^{2\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}}\mid \mathcal{Y}_T\right]^{\frac{1}{2}}\right|$$

$$=\sup_{\theta\in B(\bar\theta)}\left| E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}^2\mid \mathcal{Y}_T\right]^{\frac{1}{2}} E\left[e^{2\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}}\right]^{\frac{1}{2}}\right|.$$ 660

We know

$$\sup_{\theta\in B(\bar\theta)} E\left[\prod_{t=1}^{T}\left\{\frac{\epsilon_T(Y_t,\theta)}{\sqrt{T}}+1\right\}^2\mid \mathcal{Y}_T\right]^{\frac{1}{2}}$$

converges to a constant in $\mathbb{P}^Y$-probability and

$$\sup_{\theta\in B(\bar\theta)} E\left[e^{2\left\{Z\sigma(\theta)-\frac{\sigma^2(\theta)}{2}\right\}}\right]^{\frac{1}{2}}=\sup_{\theta\in B(\bar\theta)}\exp(\sigma(\theta)^2)^{1/2}<\infty.$$

## S3. Generalized Linear Mixed Models

### S3·1. *Exponential Families and Random Effects*

In this section we introduce a class of random effects models for which all assumptions required for 665 Theorem 1 are satisfied. We analyse the latent variable model introduced in Section 5 for the popular class of generalized linear mixed models (see e.g McCulloch & Neuhaus, 2005), where the observation density is of the form of an exponential family. We restrict attention here to the class of natural exponential family distributions, i.e. $T(y)=y$, with respect to the Lebesgue measure

$$p(y\mid\eta)=m(y)\exp\left\{\eta^{\mathrm{T}}y-A(\eta)\right\},\tag{4.1}$$

where $y$ is the natural sufficient statistic and $\eta$ denotes the natural parameter, which will be set equal to 670 the linear predictor in a generalized linear model. The function $m(y)$ is a base measure, which can be absorbed into the dominating measure. $A(\eta)$ is commonly referred to as the log-partition function and we assume that $A$ is strictly convex and increasing in $\eta$ so that the log-likelihood will be strictly concave. This assumption will be satisfied in the most common natural exponential family models including Poisson and Binomial models. In the following we will allow for multiple measurements for each group, which 675 means we have one random effect associated with multiple observations. This corresponds to the logistic mixed model of Section 7. For the conditional exponential family with $J$ repeated measurements $y_t=(y_{t,1},\dots,y_{t,J})^{\mathrm{T}}$ where $\eta_{t,j}=c_{t,j}^{\mathrm{T}}\beta+X_t$, $j=1,\dots,J$, $t=1,\dots,T$ and the random effects are centred Gaussian variables $X\sim\mathcal{N}(0,\tau^2)$ independent for each set of repeated measurements $y$. We will simplify the notation by dropping the subscript $t$ as the importance sampler for each $t$ can be considered in isolation. 680 Assume here that

$$g(y\mid x,\theta)=\prod_{j=1}^{J}m(y_j)\exp\left[\eta_j(x)y_j-A\{\eta_j(x)\}\right],\quad f(x\mid\theta)=\varphi(x;0,\tau^2),\tag{4.2}$$

where $\eta_j(x) = c_j^{\mathrm{T}}\beta + x$ and $c$ is a vector of covariates with corresponding parameter vector $\beta$. The (full) model likelihood for every observation is now given by

$$p(y, x \mid \theta) \propto \prod_{j=1}^{J} m(y_j) \exp\left[\eta_j(x) y_j - A\{\eta_j(x)\}\right] \varphi(x, 0, \tau^2).$$

Since $X$ is unobserved, we are interested in the marginal likelihood

$$\begin{aligned} p(y \mid \theta) &= \int p(y, x \mid \theta) \mathrm{d}x \\ &= \int \prod_{j=1}^{J} m(y_j) \exp\left[\eta_j(x) y_j - A\{\eta_j(x)\}\right] \varphi(x, 0, \tau^2) \mathrm{d}x. \end{aligned}$$

Consequently, the likelihood of a set of observations $y_{1:T}$, with $y_i = (y_{i,1}, \ldots, y_{i,J})$ is

$$p(y_{1:T} \mid \theta) = \prod_{t=1}^{T} \int \prod_{j=1}^{J} m(y_{t,j}) \exp\left[\eta_{t,j}(x_t) y_{t,j} - A\{\eta_{t,j}(x_t)\}\right] \varphi(x_t, 0, \tau^2) \mathrm{d}x_t.$$

We list the log-partition function as well as it's first derivative $A'(x) = \partial_x A(x)$ (which will be important later) below together with the base measure.

**Binomial.** Denote $n$ the number of trials, then

$$A(\eta) = n \log\left(1 + e^{\eta}\right), \quad A'(\eta) = \frac{n e^{\eta}}{1 + e^{\eta}}, \quad m(y) = \binom{n}{y}.$$

**Poisson.** For the Poisson family

$$A(\eta) = e^{\eta}, \quad A'(\eta) = e^{\eta}, \quad m(y) = \frac{1}{y!}.$$

## S3·2. *Asymptotic Posterior Normality*

This section establishes the Bernstein-von Mises theorem for priors having exponentially decaying tails. Denote $\Theta \subset \mathbb{R}^d$ a subset of the Euclidean space, where we take $d = 1$ without loss of generality. Consider the case of i.i.d. observations $Y_1, Y_2, \ldots$ drawn from a density $Y_i \sim f(\cdot \mid \bar{\theta})$, where $\bar{\theta} \in \Theta$ is assumed to be the "true parameter". The measure describing the distribution of the data vector $Y_{1:T} = (Y_1, \ldots, Y_T)$ is written as $P_{T,\bar{\theta}}$. Writing $\pi(\theta)$ for the prior distribution we denote the posterior density as

$$\pi_T(\theta) = \pi(\theta \mid Y_{1:T}) = \frac{\prod_{i=1}^{T} f(y_i \mid \theta)\pi(\theta)}{\int_{\Theta} \prod_{i=1}^{T} f(y_i \mid \theta)\pi(\theta)\mathrm{d}\theta}.$$

THEOREM 4. *Let the experiment be differentiable in quadratic mean at $\bar{\theta}$ with non-singular Fisher information matrix $I_{\bar{\theta}}$, and suppose that for every $\varepsilon > 0$ there exist an increasing sequence of sets $K_1 \subset K_2 \subset \ldots$ with $\cup_{i=1}^{\infty} K_i = \Theta$ with $K_T$ growing at rate $T$. Assume there exists a sequence of tests such that*

$$E(\phi_T) \to 0, \qquad \sup_{\{\|\theta - \bar{\theta}\| \geq \varepsilon\} \cap K_T} E_{\theta}^{n}(1 - \phi_T) \to 0.$$

*Furthermore, let the prior measure be absolutely continuous in a neighbourhood of $\bar{\theta}$ with a continuous positive density at $\bar{\theta}$ s.t. for $T$ large enough, we have*

$$\pi\left([-T, T]^{\complement}\right) \leq c_1 \exp(-c_2 T),$$

*where $c_1$ and $c_2$ are positive constants. Then the corresponding posterior distributions satisfy*

$$\int \left| \tilde{\pi}_T(h) - \varphi\left(h, \sqrt{T}\left(\hat{\theta}_T - \theta_0\right), I_{\bar{\theta}}^{-1}\right) \right| \mathrm{d}h \to 0 \tag{4.3}$$

*in $P_{T,\bar{\theta}}$-probability where*

$$\Delta_T(\bar{\theta}) = \frac{1}{\sqrt{T}} \sum_{i=1}^{T} \tilde{I}_{\bar{\theta}}^{-1} \frac{\partial \ell(\bar{\theta}, Y_i)}{\partial \theta}$$

*and*

$$\tilde{\pi}_T(h) = \frac{\pi_T(\bar{\theta} + h/\sqrt{T})}{T^{1/2}}$$

*is a measure on* $H = \left\{h = \sqrt{T}\left(\theta - \bar{\theta}\right) : \theta \in \Theta\right\}$.

*Proof.* The proof follows Van der Vaart (2000), Theorem 10.1, see also the lecture notes by Nickl (2012). We will show that it is enough to show convergence of the measures restricted on some arbitrarily large compact set. In order to do so, denote

$$P^C(A) = \frac{P(A \cap C)}{P(C)}$$

for any measurable set $A$ the restriction of the probability measure $P$ to the set $C$. Denote $h = \sqrt{T}\left(\theta - \bar{\theta}\right)$. We will write $P_{T,h}$ for the posterior distribution with data $Y_{1:T}$ and parameter $\bar{\theta} + h/\sqrt{T} (= \theta)$. Define the prior-weighted mixture measure over a set $C$ as

$$P_{T,C} = \int P_{T,h} \tilde{\pi}_T^C(h) \mathrm{d}h.$$

The expectation with respect to $P_{T,C}$ is calculated as

$$E_{P_{T,C}}\left\{f(Y_{1:T})\right\} = \iint f(y_{1:T}) \mathrm{d}P_{T,h}(y_{1:T}) \tilde{\pi}_T^C(h) \mathrm{d}h.$$

For any sequence of sets $A_T$ with $P_{T,\bar{\theta}}(A_T) \to 0$ it follows that $P_{T,B}(A_T) \to 0$ and vice versa, where $B$ denotes a closed ball around 0. (Two measures with this relationship are called mutually contiguous.)

This means that we can interchange convergence in probability under the measures $P_{T,B}$ and $P_{T,0}$. Let $C$ now denote a ball of size $M_T$ around 0 where $M_T \to \infty$ as $T \to \infty$. We can show that the total variation between distance between the posterior and the posterior restricted on the set $C$ vanishes by estimating

$$\left\| \tilde{\pi}_T(B) - \tilde{\pi}_T^C(B) \right\|_{\mathrm{tv}} \leq 2\tilde{\pi}_T\left(C^{\complement}\right),$$

where $\|\cdot\|_{\mathrm{tv}}$ denotes the total variation norm. We will show that the left-hand side converges to zero under $P_{T,B}$ for $B$ a closed ball around 0. We can now use the tests $\phi_T$ to bound

$$\begin{aligned}
E_{T,B}\left\{\tilde{\pi}_T(C^{\complement})\right\} &= E_{T,B}\left\{\tilde{\pi}_T\left(C^{\complement}\right)(1 - \phi_T + \phi_T)\right\} \\
&\leq E_{T,B}\left[\tilde{\pi}_T\left(C^{\complement}\right)(1 - \phi_T)\right] + E_{T,B}\left(\phi_T\right),
\end{aligned}$$

where $E_{T,B}\left(\phi_T\right) = o_{P_{T,B}}(1)$ by assumption. Now

$$\begin{aligned}
&E_{T,B}\left\{P_{H_T|Y_{1:T}}\left(C^{\complement}\right)(1 - \phi_T)\right\} \\
&= \int_B \int_{\mathbb{R}^T} \int_{C^{\complement}} (1 - \phi_T) \frac{\prod_{i=1}^T f(\bar{\theta} + g/\sqrt{T}, y_i)}{\int \prod_{i=1}^T f(\bar{\theta} + m/\sqrt{T}, y_i) \mathrm{d}\tilde{\pi}(m)} \prod_{i=1}^T f\left(\bar{\theta} + \frac{h}{\sqrt{T}}, y_i\right) \mathrm{d}y_i \frac{\mathrm{d}\tilde{\pi}(h)}{\tilde{\pi}(B)} \\
&= \frac{\tilde{\pi}(C^{\complement})}{\tilde{\pi}(B)} \int_{C^{\complement}} \int_{\mathbb{R}^T} \int_B (1 - \phi_T) \frac{\prod_{i=1}^T f(\bar{\theta} + \frac{h}{\sqrt{T}}, y_i)}{\int \prod_{i=1}^T f(\bar{\theta} + m/\sqrt{T}, y_i) \mathrm{d}\tilde{\pi}(m)} \mathrm{d}\tilde{\pi}(h) \mathrm{d}P_g^T(y) \mathrm{d}\tilde{\pi}^{C^{\complement}}(g)
\end{aligned}$$

$$= \frac{\tilde{\pi}(C^{\complement})}{\tilde{\pi}(B)} E_{T,C^{\complement}} \{\tilde{\pi}_T(B)(1-\phi_T)\}.$$

The upper bound is

$$\begin{aligned}
\frac{\tilde{\pi}(C^{\complement})}{\tilde{\pi}(B)} E^T_{C^{\complement}} \tilde{\pi}_T(B)(1-\phi_T) &= \frac{\tilde{\pi}(C^{\complement})}{\tilde{\pi}(B)} \int_{C^{\complement}} \tilde{\pi}_T(B)(1-\phi_T) P_h^T \frac{\mathrm{d}\tilde{\pi}(h \cap C^{\complement})}{\tilde{\pi}(C^{\complement})} \\
&= \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement}} \int_{\mathbb{R}^d} \tilde{\pi}_T(B)(1-\phi_T) d P_h^T(y) d\tilde{\pi}(h \cap C^{\complement}) \\
&\leq \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement}} E(1-\phi_T) \mathrm{d}\tilde{\pi}(h \cap C^{\complement}) \\
&= \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement}} E(1-\phi_T) \mathrm{d}\tilde{\pi}(h) \\
&= \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement} \cap \tilde{K}_T} E(1-\phi_T) \mathrm{d}\tilde{\pi}(h) + \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement} \cap \tilde{K}_T^{\complement}} E(1-\phi_T) \mathrm{d}\tilde{\pi}(h),
\end{aligned}$$

where $\tilde{K}_T = \{h = \sqrt{T}(\theta - \bar{\theta}) : \theta \in K_T\}$. For simplicity and without loss of generality we assume $K_T = [-T, T]$ in the following. By Van der Vaart (2000, Lemma 10.3) the tests converge exponentially fast so with $\theta = \bar{\theta} + h/\sqrt{T}$, $h = \sqrt{T}\left(\theta - \bar{\theta}\right)$, $\mathrm{d}\theta = \mathrm{d}h/\sqrt{T}$

$$\begin{aligned}
\frac{1}{\tilde{\pi}(B)} \int_{C^{\complement} \cap \tilde{K}_T} E(1-\phi_T) \mathrm{d}\tilde{\pi}(h) &= \frac{1}{\tilde{\pi}(B)} \int_{\{\|\theta-\bar{\theta}\| \geq M_T/\sqrt{T}\} \cap K_T} E_\theta\,(1-\phi_T)\,\pi(\theta) \mathrm{d}\theta \\
&= \frac{1}{\tilde{\pi}(U)} \int_{\{\|\theta-\bar{\theta}\| \geq M_T/\sqrt{T}\} \cap K_T} E_\theta\,(1-\phi_T)\,\pi(\theta) \mathrm{d}\theta \\
&= \frac{1}{\tilde{\pi}(B)} \int_{\{D' \geq \|\theta-\bar{\theta}\| \geq M_T/\sqrt{T}\} \cap K_T} E_\theta\,(1-\phi_T)\,\pi(\theta) \mathrm{d}\theta \\
&\quad + \frac{1}{\tilde{\pi}(B)} \int_{\{\|\theta-\bar{\theta}\| \geq D'\} \cap K_T} E_\theta\,(1-\phi_T)\,\pi(\theta) \mathrm{d}\theta \\
&= c_2 \int_{\{D' \geq \|\theta-\bar{\theta}\| \geq M_T/\sqrt{T}\} \cap K_T} \exp\left(-DT \left\|\theta - \bar{\theta}\right\|^2\right) \mathrm{d}\theta \\
&\quad + \frac{1}{\tilde{\pi}(B)} \int_{\{\|\theta-\bar{\theta}\| \geq D'\} \cap K_T} \exp(-c_3 T) \pi(\theta) \mathrm{d}\theta \\
&\leq c_2 \int_{\{h : h \geq M_T\} \cap K_T} \exp\left(-DT \left\|\theta - \bar{\theta}\right\|^2\right) T^{1/2} \mathrm{d}\theta \\
&\quad + 2 c_3 T^{1/2} \exp(-c_4 T),
\end{aligned}$$

where we used $\tilde{\pi}(B) \geq 1/(c_3 T^{1/2})$ for some constant $c_3$ because the prior is positive and continuous at $\bar{\theta}$. For the second part

$$\begin{aligned}
\frac{1}{\tilde{\pi}(B)} \int_{C^{\complement} \cap \tilde{K}_T^c} E_h(1-\phi_T) d\tilde{\pi}(h) &\leq \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement} \cap \tilde{K}_T^{\complement}} \mathrm{d}\tilde{\pi}(h) \\
&= \frac{1}{\tilde{\pi}(B)} \int_{C^{\complement} \cap \tilde{K}_T^{\complement}} \pi(\bar{\theta} + h/\sqrt{T}) \mathrm{d}h \\
&\leq c_3 T^{1/2} \int_{K_T^{\complement}} \pi(\theta) \mathrm{d}\theta \\
&\leq c_3 T^{1/2} \cdot c_1 \exp(-c_2 T).
\end{aligned}$$

As $T \to \infty$ we have

$$\|\tilde{\pi}_T - \tilde{\pi}_T^C\|_{\mathrm{tv}} \to 0$$

in $P_{T,B}$-probability and by contiguity also in $P_{T,\bar{\theta}}$.

Similarly, for a Gaussian distribution with means $\sup|\mu_T| < \infty$ and variance $\sigma^2$ we have

$$\left\|\mathcal{N}\left(\mu_T, \sigma^2\right) - \mathcal{N}^C\left(\mu_T, \sigma^2\right)\right\| \leq 2\mathcal{N}\left(\mu_T, \sigma^2\right)\left(C^{\complement}\right).$$

We know that $\Delta_{T,\bar{\theta}}$ is uniformly tight, i.e. for any $\varepsilon > 0$ there exists $K$ such that $\sup_T P\left(\left|\Delta_{T,\bar{\theta}}\right| \leq K\right) = 1 - \varepsilon$. Hence, with probability $1 - \varepsilon$

$$\left\|\mathcal{N}\left(\Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right) - \mathcal{N}^C\left(\Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)\right\| \leq 2\mathcal{N}\left(\Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)(C^{\complement})$$

by choosing $M$ (the radius of $C$) sufficiently large. Hence, by the triangle inequality we have to show that

$$\int \left|\tilde{\pi}_T^C(h) - \varphi^C\left(h; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)\right| \mathrm{d}h \to 0$$

in $P_{T,0}$-probability. Denoting $x^+ = \max\{0, x\}$

$$\begin{aligned}
&\frac{1}{2}\int \left|\tilde{\pi}_T^C(h) - \varphi^C\left(h; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)\right| \mathrm{d}h \\
&= \int \left(1 - \frac{\varphi^C\left(h; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)}{\tilde{\pi}_T^C(h)}\right)^+ \tilde{\pi}_T^C(h)\mathrm{d}h \\
&= \int \left(1 - \frac{\varphi^C\left(h; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)\int 1_C f_{T,g}^C(g)\pi(g)\mathrm{d}g}{1_C\tilde{\pi}_T^C(h)}\right)^+ \tilde{\pi}_T^C(h)\mathrm{d}h \\
&= \int \left(1 - \int \frac{1_C(g) f_{T,g}^C(g)\pi(g)\varphi^C\left(h; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)}{1_C(h) f_{T,h}^C(h)\pi(h)\varphi^C\left(g; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)}\varphi^C\left(g; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)\mathrm{d}g\right)^+ \tilde{\pi}_T^C(h)\mathrm{d}h \\
&\leq \iint \left(1 - \frac{f_{T,g}^C(g)\pi(g)\varphi^C\left(h; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)}{f_{T,h}^C(h)\pi(h)\varphi^C\left(g; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)}\right)^+ \varphi^C\left(g; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)\mathrm{d}g\tilde{\pi}_n^C(h)\mathrm{d}h \\
&\leq \left\{\sup_{x\in C}\varphi^C\left(x; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)\right\}\iint \left(1 - \frac{f_{T,g}^C(g)\pi(g)\varphi^C\left(h; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)}{f_{T,h}^C(h)\pi(h)\varphi^C\left(g; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)}\right)^+ \mathrm{d}g\tilde{\pi}_T^C(h)\mathrm{d}h.
\end{aligned}$$

By dominated convergence it is enough to conclude that this quantity goes to 0 in

$$\begin{aligned}
P_{T,C}(\mathrm{d}y)\tilde{\pi}_T^C(\mathrm{d}h)\lambda_C(\mathrm{d}g) &= \int P_{T,x}(\mathrm{d}y)\tilde{\pi}_T^C(h)\mathrm{d}h\lambda_C(\mathrm{d}g) \\
&= \int \prod_{i=1}^T f\left(\theta + s/\sqrt{T}, y_i\right)\frac{\prod_{i=1}^T f\left(\theta + h/\sqrt{T}, y_i\right)\tilde{\pi}^C(h)\mathrm{d}h}{\int\prod_{i=1}^T f\left(\theta + u/\sqrt{T}, y_i\right)\tilde{\pi}^C(u)\mathrm{d}u}\mathrm{d}s\lambda_C(\mathrm{d}g) \\
&= \prod_{i=1}^T f\left(\theta + h/\sqrt{T}, y_i\right)\tilde{\pi}^C(h)\mathrm{d}h\lambda_C(\mathrm{d}g) \\
&= P_{T,C}(\mathrm{d}y)\tilde{\pi}^C(h)\mathrm{d}h\lambda_C(\mathrm{d}g)
\end{aligned}$$

probability. Under Theorem 7.2 in Van der Vaart (2000) mean-square differentiability of the likelihood implies that the likelihood ratio allows for the LAN (Van der Vaart, 2000, Definition 7.14) expansion

$$\frac{\prod_{i=1}^T f(\theta + g/\sqrt{T}, y_i)}{\prod_{i=1}^T f(\theta + h/\sqrt{T}, y_i)} =$$
$$= \prod_{i=1}^T \frac{f(\theta + g/\sqrt{T}, y_i)}{f(\theta, y_i)} \Big/ \prod_{i=1}^T \frac{f(\theta + h/\sqrt{T}, y_i)}{f(\theta, y_i)}$$
$$= \exp\left( \frac{1}{\sqrt{T}} \sum_{i=1}^T g^T \ell'_\theta(y_i) - \frac{1}{2} g^T I_\theta g - \frac{1}{\sqrt{T}} \sum_{i=1}^T h^T \ell'_\theta(y_i) - \frac{1}{2} h^T I_\theta h + o_{P_\theta}(1) \right)$$

and thus as $T \to \infty$ and using continuity of the prior $\pi$ at $\bar{\theta}$ we have

$$1 - \frac{f^C_{T,g}(g)\pi(g)\varphi^C\left(h; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)}{f^C_{T,h}(h)\pi(h)\varphi^C\left(g; \Delta_{T,\bar{\theta}}, I_{\bar{\theta}}^{-1}\right)} \to 0$$

which yields the result. □

*Remark* 3. *i)* The centring sequence $\Delta_{T,\theta}$ can be replaced by any best regular estimator. To see this note that following Van der Vaart (2000, Theorem 8.14) any best regular estimator, $\hat{\theta}_T$, satisfies the expansion

$$\sqrt{T}(\hat{\theta}_T - \bar{\theta}) = \frac{1}{\sqrt{T}} \sum_{i=1}^T \tilde{I}_{\bar{\theta}}^{-1} \frac{\partial \ell(\bar{\theta}, Y_i)}{\partial \theta} + o_{P_{T,\bar{\theta}}}(1)$$

and thus

$$\Delta_T(\bar{\theta}) - \sqrt{T}\left(\hat{\theta}_T - \bar{\theta}\right) \to 0$$

in $P_{T,\theta_0}$-probability as $T \to \infty$. Since

$$\left\| \mathcal{N}\left(\Delta_{T,\bar{\theta}}, \tilde{I}_{\bar{\theta}}^{-1}\right) - \mathcal{N}\left\{\sqrt{T}\left(\hat{\theta}_T - \bar{\theta}\right), \tilde{I}_{\bar{\theta}}^{-1}\right\} \right\| \lesssim \left\| \sqrt{T}\left(\hat{\theta}_T - \bar{\theta}\right) - \Delta_{T,\bar{\theta}} \right\| \to 0$$

in probability.

*ii)* Under regularity conditions (Van der Vaart, 2000, Theorem 5.39) the maximum likelihood estimator is best regular and can be used as a centring sequence following the argument in *i)*.

We will now apply this Bernstein-von Mises result to our exponential family models. Hence, consider again the likelihood contribution of every observation $y$,

$$p(y \mid \beta, \tau) = \int \prod_{j=1}^J m(y_j) \exp\left\{(c_j^{\mathrm{T}}\beta + x)y_j - A(c_j^{\mathrm{T}}\beta + x)\right\} \varphi(x, 0, \tau^2)\mathrm{d}x. \tag{4.4}$$

For simplicity we assume that the exogenous variables $c_j$ are all identical and that $\Theta$ is a subset of $\mathbb{R}$. Let $A$ be continuously differentiable (e.g. the Binomial and Poisson models introduced above). The prior can be easily chosen to fulfil the conditions of the updated Bernstein–von Mises theorem. The other conditions need further analysis. In order to show differentiability in quadratic mean it is sufficient to prove that the map $\theta \mapsto p(y \mid \theta)^{1/2}$ is continuously differentiable. By Lemma 7.6 in Van der Vaart (2000) we need to show that

$$\theta \mapsto p(y \mid \theta)^{1/2} = \left[ \int m(y) \exp\left\{(c^{\mathrm{T}}\beta + x) \cdot y - A(c^{\mathrm{T}}\beta + x)\right\} \varphi(x, 0, \tau^2)\mathrm{d}x \right]^{1/2}$$

is continuously differentiable for all $y$. Firstly,

$$\frac{\partial}{\partial \theta} p(y \mid \theta)^{1/2} = \frac{1}{2p(y \mid \theta)^{1/2}} \partial_\theta p(y \mid \theta).$$

It is easy to see that $\theta \mapsto p(y \mid \theta)$ and $\theta \mapsto \partial_\theta p(y \mid \theta)$ are continuous. The fisher information is well defined, continuous in $\theta$ and positive since

$$\begin{aligned} I_\theta &= E\left[\{\partial_\theta \log p(Y \mid \theta)\}^2\right] \\ &= \int \{\partial_\theta \log p(y \mid \theta)\}^2 \, p(y \mid \theta) \mathrm{d}y > 0 \end{aligned}$$

whenever $\partial_\theta \log p(y \mid \theta)$ is not identically 0 for all $y$. The multivariate case is more involved and treated for example in Mukerjee & Sutradhar (2002) for the Binomial and Poisson case. In order to ensure the existence of the tests consider $K_1 \subset K_2 \subset \ldots$ an increasing sequence of compact sets with $\cup_{i=1}^{\infty} K_i = \Theta$. Then, if the model is identifiable and continuous in total variation norm, Lemma 10.6 in Van der Vaart (2000), and a diagonal argument similar to that in the proof of (Van der Vaart, 2000, Lemma 10.6), ensures the existence of a sequence of estimators $\hat{\theta}_T$ such that $\sup_{\theta \in K_T} P_\theta(|\hat{\theta}_T - \theta| \geq \varepsilon) \to 0$ whence we have, see for example (Nickl, 2012, Lemmas 1,2 in Section 2.2.3),

$$E_{\bar{\theta}}(\phi_T) \to 0, \qquad \sup_{\{\|\theta - \bar{\theta}\| \geq \varepsilon\} \cap K_T} E_{T,\theta}\,(1 - \phi_T) \to 0.$$

Since our model has a density with respect to the Lebesgue measure continuity in total variation is trivially the case as we can write the total variation distance as

$$\|P_\theta - P_{\theta'}\|_{\mathrm{tv}} = \int \left| p(y \mid \theta) - p(y \mid \theta') \right| \mathrm{d}y.$$

Therefore, by Scheffé's lemma, continuity in the parameter already implies convergence of the integral and therefore continuity in the total variation distance. To conclude that our models are indeed identifiable it is enough to ensure that

*i*) the integral

$$E(Y) = E\left[A'(k + X)\right] = \int_{\mathbb{R}} A'(k + \tau x) \varphi(x; 0, 1) \mathrm{d}x < \infty,$$

for all $k$, $\tau$ and

*ii*) the equation

$$\frac{A'(c^T \beta_1 + \tau_1 x)}{\tau_1} = \frac{A'(c^T \beta_2 + \tau_2 x)}{\tau_2} \quad \text{for all } c \text{ and } x$$

has no solution,

see Labouriau (2014). These conditions are fulfilled for the Binomial case, $A'(\eta) = ne^\eta/(1 + e^\eta)$, and Poisson case $A'(\eta) = e^\eta$.

### S3·3. *Importance Sampling with Univariate Random Effects*

We will now consider Assumption 3 in the context of generalized linear mixed models, which we will prove using Assumption 4 and Theorem 3. In the following we will first consider a univariate random effect and a Gaussian importance sampling proposal. This will include the example of Section 7. In addition we will show how fatter tails in the proposal affect the existence of moments by considering a univariate $t$-proposal. Recall that we are interested in bounds on

$$E^Y\left[\sup_{\theta \in B(\overline{\theta})} E^{X|Y}\left\{\overline{w}(Y, X, \theta)^a\right\}\right] = E^Y\left[\sup_{\theta \in B(\overline{\theta})} \frac{E^{X|Y}\{w(Y, X, \theta)^a\}}{p(Y \mid \theta)^a}\right], \tag{4.5}$$

where $a > 0, \theta = (\beta, \tau)$ and $B(\bar{\theta}) \subset \Theta$ denotes a closed $\varepsilon$-ball around $\bar{\theta}$. For additional clarity, we write $E^Y$ and $E^{X|Y}$ for the expectations over $Y$ and $X$ given $Y$, respectively. Consider the Gaussian proposal centred at the mode

$$q(x \mid y) = \varphi(x; \widehat{x}, \tau_q^2), \tag{4.6}$$

where $\tau_q^2$ denotes the proposal variance and $\widehat{x}$ is the mode of $h(x; y) = g(y \mid x) f(x)$ and fulfils the first order condition

$$\widehat{x} = \tau^2 \left\{ S - \widetilde{A}'(\widehat{x}) \right\}, \tag{4.7}$$

where $\widetilde{A}'(x) = \sum_{j=1}^J A'(c_j^{\mathrm{T}}\beta + x)$ with $A'(z) = \partial_z A(z)$ and $S = \sum_{j=1}^J y_j$. For later convenience we define the unnormalized proposal density

$$\widetilde{q}(x; y) = \frac{q(x \mid y)}{q(\widehat{x} \mid y)},$$

where $q(x \mid y)$ is the proposal density. For a symmetric proposal distribution centred at $\widehat{x}$ the term $q(\widehat{x} \mid y)$ is simply an inverse normalizing constant, which only involves the proposal parameters. For the Gaussian proposal

$$\widetilde{q}(x; y) = \exp\left\{ -\frac{(x - \widehat{x})^2}{2\tau^2} \right\}, \quad q(\widehat{x} \mid y) = \frac{1}{(2\pi \tau_q^2)^{1/2}}. \tag{4.8}$$

Associated with this we introduce the modified weight which is defined as

$$\widetilde{w}(x, y) = \frac{g(y \mid x) f(x)}{g(y \mid \widehat{x}) f(\widehat{x})} \frac{1}{\widetilde{q}(x; y)} = \frac{h(x; y)}{h(\widehat{x}; y)} \frac{1}{\widetilde{q}(x; y)}, \tag{4.9}$$

where $h(x; y) = g(y \mid x) f(x)$. These weights are easier to work with as $\widetilde{w}(x, y) = 1$ when $x = \widehat{x}$. It is easily seen that

$$\widetilde{w}(x, y) = \frac{h(x; y)}{q(x \mid y)} \frac{q(\widehat{x} \mid y)}{h(\widehat{x}; y)} = w(x, y) \frac{q(\widehat{x} \mid y)}{h(\widehat{x}; y)},$$

so that the modified weights are proportional to the standard weights $w(x, y)$ as a function of $x$. We can recast the expectation (4.5) as

$$\begin{aligned} E^Y \left[ \sup_{\theta \in B(\overline{\theta})} E^{X|Y} \left\{ \overline{w}(Y, X, \theta)^a \right\} \right] &= E^Y \left[ \sup_{\theta \in B(\overline{\theta})} \frac{E^{X|Y} \{ w(X, Y, \theta)^a \}}{p(Y \mid \theta)^a} \right] \\ &= E^Y \left[ \sup_{\theta \in B(\overline{\theta})} \frac{E^{X|Y} \{ \widetilde{w}(X, Y, \theta)^a \}}{E^{X|Y} \{ \widetilde{w}(X, Y, \theta) \}^a} \right]. \end{aligned} \tag{4.10}$$

The log-density of the observations is given by

$$\begin{aligned} \log g(y \mid x) &= \sum_{j=1}^J \{ \log m(y_j) + y_j \eta_j - A(\eta_j) \} \\ &= \sum_{j=1}^J \{ \log m(y_j) + y_j c_j^{\mathrm{T}} \beta + y_j x - A(c_j^{\mathrm{T}} \beta + x) \} \\ &= k(y) + x(J \overline{y}) - \widetilde{A}(x), \end{aligned}$$

where $k(y)$ represents constant values (which do not depend upon $x$), $J\overline{y} = \sum_{j=1}^{J} y_j$ and $\widetilde{A}(x) = \sum_{j=1}^{J} A(c_j^{\mathrm{T}}\beta + x)$. Hence, we get

$$\log h(x; y) = \log g(y \mid x) f(x) = c + x(J\overline{y}) - \widetilde{A}(x) - \frac{x^2}{2\tau^2}. \tag{4.11}$$

We will proceed by deriving bounds for the denominator and enumerator of (4.5) separately. We present the following lemma on the denominator without reference to the Gaussian proposal, because it holds for general proposal distribution.

LEMMA 10. *Consider the exponential family model with repeated measurement $j = 1, \ldots, J$ and Gaussian random effects. For general proposal density $q(x \mid y)$ we have*

$$\frac{1}{E^X\{\widetilde{w}(X, y)\}} \leq \frac{(2\pi)^{1/2}}{C}(b+1),$$

*where $b = \tau\widetilde{A}'(\widehat{x})$ and $C = q(\widehat{x} \mid y)(2\pi\tau^2)^{1/2}$.*

*Proof of Lemma* 10. For given observation $y$, the expectation of the rescaled weights is

$$\begin{aligned} E^X\{\widetilde{w}(X, y)\} &= \int \widetilde{w}(x, y) q(x \mid y)\mathrm{d}x \\ &= \int \frac{h(x; y)}{h(\widehat{x}; y)} \frac{q(x \mid y)}{\widetilde{q}(x; y)}\mathrm{d}x \\ &= q(\widehat{x} \mid y) \int \frac{h(x; y)}{h(\widehat{x}; y)}\mathrm{d}x. \end{aligned}$$

Write again $S = J\bar{y} = \sum_{j=1}^{J} y_j$. Since $\widetilde{A}$ is an increasing function we obtain for $x \leq \widehat{x}$,

$$\begin{aligned} \log h(x; y) - \log h(\widehat{x}; y) &= -\widetilde{A}(x) + xS - \frac{1}{2}\frac{x^2}{\tau^2} + \widetilde{A}(\widehat{x}) - \widehat{x}S + \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} \\ &\geq (x - \widehat{x})S - \frac{1}{2}\frac{x^2}{\tau^2} + \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} \\ &= R_2 - \frac{1}{2}\frac{\{x - \tau^2 S\}^2}{\tau^2}, \end{aligned}$$

where

$$R_2 = \frac{\tau^2 S^2}{2} - \widehat{x}S + \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} = \frac{\tau^2\widetilde{A}'(\widehat{x})^2}{2},$$

by using the first order condition for the mode $\widehat{x} = \tau^2\{S - \widetilde{A}'(\widehat{x})\}$. Therefore

$$E^X\{\widetilde{w}(X, y)\} \geq q(\widehat{x} \mid y)\left(2\pi\tau^2\right)^{1/2} \exp\left\{\frac{\tau^2\widetilde{A}'(\widehat{x})^2}{2}\right\} \Phi\{-\tau\widetilde{A}'(\widehat{x})\}.$$

Consider the inequality due to Birnbaum (1942)

$$\frac{\exp(-b^2/2)}{1 - \Phi(b)} < \left(\frac{\pi}{2}\right)^{1/2}\left\{b + (b^2+4)^{1/2}\right\}$$

Setting $b = \tau\widetilde{A}'(\widehat{x})$ and $C = q(\widehat{x} \mid y)(2\pi\tau^2)^{1/2}$ gives

$$\begin{aligned} \frac{1}{E^X\{\widetilde{w}(X, y)\}} &\leq \frac{\exp(-b^2/2)}{C\{1 - \Phi(b)\}} \\ &\leq C^{-1}\left(\frac{\pi}{2}\right)^{1/2}\left\{b + (b^2+4)^{1/2}\right\} \end{aligned}$$

$$\leq \frac{(2\pi)^{1/2}}{C}(b+1),$$

as $(b^2+4)^{1/2} \leq b+2$. □

Using Lemma 10 we can set

$$\sup_{\theta \in B(\overline{\theta})} \frac{1}{E^X\left\{\widetilde{w}(X,y,\theta)\right\}} \leq \sup_{\theta \in B(\overline{\theta})} \frac{1}{q(\widehat{x} \mid y)\tau}\left\{\tau \widetilde{A}'(\widehat{x})+1\right\}.$$

We will use this result in the following corollary.

COROLLARY 1. *Assume one of the following condition holds:*

*(i)* $\sup_x \widetilde{A}'(x) < \infty$,
*(ii)* $E^Y(Y^a) < \infty$ *and* $\sup_{\theta \in B(\bar{\theta})} \widetilde{A}'(0) < \infty$.

*Then taking the expectation over $Y$, we have*

$$E^Y\left[\sup_{\theta \in B(\overline{\theta})} \frac{1}{E^{X|Y}\{\widetilde{w}(X,Y)\}^a}\right] < \infty.$$

*Proof.* Applying Lemma 10 with $b = \tau\widetilde{A}'(\widehat{x})$ yields

$$E^Y\left[\sup_{\theta \in B(\overline{\theta})} \frac{1}{E^X\{\widetilde{w}(X,Y)\}^a}\right] \leq E^Y\left[\sup_{\theta \in B(\overline{\theta})}\left\{\frac{\tau\widetilde{A}'(\widehat{x})+1}{C\tau}\right\}^a\right], \tag{4.12}$$

where we write $C = q(\widehat{x} \mid y)$ which only involved parameters of the proposal distribution. The right-hand side of (4.12) is finite provided $E^Y\{\sup_{\theta \in B(\overline{\theta})} \widetilde{A}'(\widehat{x})^a\} < \infty$. This concludes the proof for *(i)*. For *(ii)* we need to control the function $\widetilde{A}'(\widehat{x})$. Therefore, it is useful to establish the behaviour of $\widetilde{A}'(\widehat{x})$ in terms of the random variables $y = (y_1,\ldots,y_J)$. Recall the first order condition (4.7)

$$\widehat{x} = \tau^2\{J\overline{y} - \widetilde{A}'(\widehat{x})\},$$

where the sufficient statistic is $S = J\overline{y} = \sum_{j=1}^J y_j$. It is easily established that $\widetilde{A}'(\widehat{x}) \leq \max\{\widetilde{A}'(0), J\overline{y}\}$. To see this note

$$\partial_x \log h(x;y) = J\overline{y} - \widetilde{A}'(x) - \frac{x}{\tau^2}. \tag{4.13}$$

The function $\widetilde{A}'(x)$ is monotonically increasing. If $\widetilde{A}'(0) < J\overline{y}$, then at $x=0$, $\partial_x \log h(x;y) > 0$ and at $x = \widetilde{x}$, where $\widetilde{A}'(\widetilde{x}) = J\overline{y}$, $\partial_x \log h(x;y) < 0$ since $\widetilde{x} > 0$. Similarly, if $\widetilde{A}'(0) < J\overline{y}$ then at $x=0$, $\partial_x \log h(x;y) < 0$ and at $x = \widetilde{x}$, $\partial_x \log h(x;y) > 0$. As a consequence, the mode of the concave function $\log h(x;y)$, $\widehat{x}$ is always between 0 and $\widetilde{x}$, where $\widetilde{A}'(\widetilde{x}) = J\overline{y}$. This yields $\widetilde{A}'(\widehat{x}) \leq \max\{\widetilde{A}'(0), J\overline{y}\}$ so that

$$\begin{aligned}
E^Y\left\{\sup_{\theta \in B(\overline{\theta})} \widetilde{A}'(\widehat{x})^a\right\} &\leq E^Y\left[\sup_{\theta \in B(\overline{\theta})} \max\{\widetilde{A}'(0), S\}^a\right] \\
&\leq E^Y\left[\max\left\{\sup_{\theta \in B(\overline{\theta})} \widetilde{A}'(0), S\right\}^a\right] \\
&= \sup_{\theta \in B(\overline{\theta})} \widetilde{A}'(0)^a \mathbb{P}^Y\left\{S < \sup_{\theta \in B(\overline{\theta})} \widetilde{A}'(0)\right\} + \int_{\sup_{\theta \in B(\overline{\theta})} \widetilde{A}'(0)}^{\infty} s^a \mathrm{d}F_S(s).
\end{aligned}$$

The last quantity is finite whenever $\sup_{\theta \in B(\overline{\theta})} \widetilde{A}'(0) < \infty$ and $E^Y(Y^a) < \infty$. □

*Remark* 4 (*Examples with Gaussian proposal*). If the proposal is a Gaussian centred at the mode $q(x \mid y) = \varphi(x;\widehat{x},\tau_q^2)$ and $C$ as defined in Lemma 10, then $C = \tau/\tau_q$. For the Binomial case, we know that $\sup_x \widetilde{A}'(x) < \infty$ and therefore condition $(i)$ of the preceding Corollary 1 is fulfilled. For the Poisson case $\widetilde{A}'(x)$ is not bounded, but we can use the second part of the corollary. Note that $\widetilde{A}'(x)$ is continuous and therefore $\widetilde{A}'(0)$ can be bounded in a neighbourhood small enough. In addition, if the Poisson model is true, it is straightforward to establish that the moments $E(Y^a)$ exist for all $a > 0$ and we can therefore conclude by part *(ii)*.

Having established conditions to ensure

$$E^Y \left[ \sup_{\theta \in B(\overline{\theta})} E^{X|Y} \{\widetilde{w}(X, Y)\}^{-a} \right] < \infty$$

we can bound (4.10) whenever there exists a constant $K < \infty$ such that

$$\sup_{y \in \mathsf{Y}} \sup_{\theta \in B(\overline{\theta})} E^{X|Y} \left\{ \widetilde{w}(X, y)^a \right\} < K.$$

In the following we will provide conditions for Gaussian and $t$-distributed proposals.

PROPOSITION 5. *Consider the Gaussian proposal* (4.6) *and some exponent* $a > 0$. *Then*

$$E^X\{\widetilde{w}(X, y)^a\} < \infty$$

*if and only if* $\tau_q^2 > \frac{(a-1)}{a}\tau^2$, *where* $\tau^2$ *is the variance of the random effects term. If this condition is satisfied then*

$$E^X\{\widetilde{w}(X, y)^a\} \leq \left\{ \frac{a\tau_q^2 - (a-1)\tau^2}{\tau^2} \right\}^{-\frac{1}{2}},$$

*independent of* $y$.

*Proof.* For brevity we define the sum $S = J\overline{y} = \sum_{j=1}^J y_j$ and again have $\widetilde{A}(x) = \sum_{j=1}^J A(cj^{\mathrm{T}}\beta + x)$. Note that $x \mapsto \widetilde{A}(x)$ is convex and thus always dominates its chord

$$\widetilde{A}(x) \geq \widetilde{A}(\widehat{x}) + \widetilde{A}'(\widehat{x})(x - \widehat{x})$$

for any values $x, \widehat{x}$. Then the modified proposal form $\widetilde{q}(x; y)$ is given by (4.8), so

$$\begin{aligned}
\log \widetilde{w}(x, y) &= \log h(x; y) - \log h(\widehat{x}; y) - \log \widetilde{q}(x; y) \\
&= xS - \widetilde{A}(x) - \frac{1}{2}\frac{x^2}{\tau^2} \\
&\quad - \widehat{x}S + \widetilde{A}(\widehat{x}) + \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} + \frac{1}{2}\frac{(x - \widehat{x})^2}{\tau_q^2}.
\end{aligned}$$

This is, by design, zero at $x = \widehat{x}$ and can be bounded as

$$\begin{aligned}
\log \widetilde{w}(x, y) &\leq \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} + \{S - \widetilde{A}'(\widehat{x})\}(x - \widehat{x}) - \frac{1}{2}\frac{x^2}{\tau^2} + \frac{1}{2}\frac{(x - \widehat{x})^2}{\tau_q^2} \\
&= \frac{1}{2}(x - \widehat{x})^2 d,
\end{aligned}$$

by noting the first order condition that $\widehat{x}/\tau^2 = S - \widetilde{A}'(\widehat{x})$. The constant $d$ is defined to be

$$d = \frac{1}{\tau_q^2} - \frac{1}{\tau^2},$$

and $d > 0$ if we choose $\tau_q^2 < \tau^2$. Hence

$$E^X\{\widetilde{w}(X, y)^a\} \leq E^X \left[ \exp\left\{ \frac{ad}{2}(X - \widehat{x})^2 \right\} \right], \tag{4.14}$$

where again the expectation is with respect to $q(x \mid y) = \varphi(x \mid \widehat{x}, \tau_q^2)$. As $a > 0$, clearly the above expectation exists if $d \leq 0$ which would imply choosing $\tau_q^2 \geq \tau^2$. To obtain a precise condition we note that

$$\frac{(X - \widehat{x})^2}{\tau_q^2} \sim \chi_1^2.$$

Considering the moment generating function of the $\chi^2$-distribution we know that the expectation (4.14) exists provided

$$ad\tau_q^2 = a\left(1 - \frac{\tau_q^2}{\tau^2}\right) < 1, \quad \text{i.e.} \quad \tau_q^2 > \frac{(a-1)}{a}\tau^2. \tag{4.15}$$

If this inequality holds, the moment generating function of the $\chi^2$-distribution exists and we have

$$E^X\left[\exp\left\{\frac{ad}{2}(X - \widehat{x})^2\right\}\right] = \left(1 - ad\tau_q^2\right)^{-1/2}.$$

Finally we obtain

$$E^X\{\widetilde{w}(X, y)^a\} \leq \left\{1 - a\left(1 - \frac{\tau_q^2}{\tau^2}\right)\right\}^{-1/2} = \left\{\frac{a\tau_q^2 - (a-1)\tau^2}{\tau^2}\right\}^{-1/2}. \tag{4.16}$$

as required. □

Note that by the upper bound in Proposition 5 still depends on parameters via the variance term $\tau$. However, since the dependence is continuous we can find an upper bound over any compact set. Thus, we have the simple corollary.

COROLLARY 2. *Under the conditions of Proposition* 5 *there exists a constant* $K_1 < \infty$ *such that*

$$\sup_{\theta \in B(\overline{\theta})} E^X\{\widetilde{w}(X, y, \theta)^a\} \leq K_1$$

*independent of* $y$.

We can summarize the results so far in the following theorem.

THEOREM 5. *Consider the random effects model* (4.2) *and assume we have an importance sampling estimator with proposal distribution*

$$q(x \mid y) = \varphi(x; \widehat{x}, \tau_q^2)$$

*and proposal variance* $\tau_q^2 > \frac{(a-1)}{a}\tau^2$. *Assume additionally that either*

*i*) $\sup_x \widetilde{A}'(x) < \infty$ *or*
*ii*) $E^Y(Y^a) < \infty$ *and* $\sup_{\theta \in B(\bar{\theta})} \widetilde{A}'(0) < \infty$.

*Then*

$$E^Y\left[\sup_{\theta \in B(\overline{\theta})} E^X\left\{\overline{w}(Y, X, \theta)^a\right\}\right] < \infty.$$

*Proof.* We have

$$E^Y\left[\sup_{\theta \in B(\overline{\theta})} E^X\left\{\overline{w}(Y, X, \theta)^a\right\}\right] = E^Y\left[\sup_{\theta \in B(\overline{\theta})} \frac{E^X\{\widetilde{w}(X, Y, \theta)^a\}}{E^X\{\widetilde{w}(X, Y, \theta)\}^a}\right]$$

$$\leq E^Y\left[\sup_{\theta\in B(\overline{\theta})} \frac{K_1}{E^X\left\{\widetilde{w}(X,Y,\theta)\right\}^a}\right] < \infty.$$

where the first inequality is by Corollary 2 and the second by Corollary 1. □

For the logistic model of Section 7, $\widetilde{A}'(x)$ is bounded above by a constant. Indeed

$$\widetilde{A}'(x) = \sum_{j=1}^{J} A'(c_j^{\mathrm{T}}\beta + x) = \sum_{j=1}^{J} \frac{e^{c_j^{\mathrm{T}}\beta+x}}{1+e^{c_j^{\mathrm{T}}\beta+x}}.$$

Hence, we know (see Remark 4) that

$$E^Y\left[\sup_{\theta\in B(\overline{\theta})} E^{X|Y}\left\{\overline{w}(X,Y)^a\right\}\right] < \infty$$

for all $a$ if we take, for example, $\tau_q^2 = \tau^2$. We note, however, that the proposal may not be particularly efficient as the proposal variance would ideally be made to be proportional to $1/J$, where $J$ represents the number of observations associated with each latent variate. Hence, taking $\tau_q^2 = \tau^2$, for example, may be much too large as a choice for $\tau_q^2$. This naturally leads to consideration of the $t$-distribution which has heavier tails, see for example (Owen, 2013, Chapter 9) and so controls the numerator term. We consider the $t$-distribution proposal centred at the mode, with scaling $\tau_q^2$, so that $q(x \mid y) = t_\nu(x \mid \widehat{x}, \tau_q^2)$. For the $t$-proposal, we have

$$\widetilde{q}(x;y) = \left\{1 + \frac{(x-\widehat{x})^2}{\nu\tau_q^2}\right\}^{-(\nu+1)/2}, \quad q(\widehat{x} \mid y) = \frac{\sqrt{\nu\pi}\,\Gamma(\nu/2)\,\tau_q}{\Gamma\{(\nu+1)/2\}}. \tag{4.17}$$

We proceed in the same manner as in the Gaussian case. First we compute the bound from Lemma 10 for the $t$-distribution. Assume the proposal is a $t$-distribution centred at the mode $q(x \mid y,\theta) = t_\nu(x \mid \widehat{x}, \tau_q^2)$, then

$$C = \frac{\tau}{\tau_q}\sqrt{\frac{2}{\nu}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}$$

and thus

$$\frac{1}{E^X\left\{\widetilde{w}(X,y)\right\}} \leq \frac{\tau_q}{\tau}\sqrt{\frac{\nu}{2}}\frac{\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu+1}{2}\right)}(b+1).$$

PROPOSITION 6. *For the target $h(x;y)$ of (4.11) with $q(x \mid y,\theta) = t_\nu(x \mid \widehat{x}, \tau_q^2)$ specified above we shall assume that the function $x \mapsto A(x)$ is a monotonically non-decreasing convex function. Then,*

$$E^{X|Y}\left\{\widetilde{w}(X,Y)^a\right\} \leq K_2^a,$$

*where*

$$K_2 = \left\{\frac{\tau^2}{\tau_q^2}\frac{(\nu+1)}{\nu}\right\}^{\frac{(\nu+1)}{2}} \exp\left\{\frac{\nu}{2}\left(\frac{\tau_q^2}{\tau^2} - 1 - \frac{1}{\nu}\right)\right\},$$

*for $\tau_q^2 < \frac{(\nu+1)}{\nu}\tau^2$ and $K_2 = 1$ for $\tau_q^2 \geq \frac{(\nu+1)}{\nu}\tau^2$.*

Unlike the Gaussian proposal above, the $t$-distributed proposal does not have any restriction on how small the variance $\tau_q^2$ can be. This might be chosen, for example, according to the second derivative of $\log h(x;y)$ at 0 so that $\tau_q^{-2} = \tau^{-2} + \widetilde{A}''(0)$. This would reflect the influence of a large number of repeated observations, $J$.

*Proof of Proposition* 6. Recall that $x \mapsto A(x)$ is convex and thus always dominates its chord

$$\widetilde{A}(x) \geq \widetilde{A}(\widehat{x}) + \widetilde{A}'(\widehat{x})(x - \widehat{x})$$

for any values $x, \widehat{x}$. For the modified log weight this yields

$$\begin{aligned}\log \widetilde{\omega}(x, y) &= \log h(x; y) - \log h(\widehat{x}; y) - \log \widetilde{q}(x; y) \\ &= xS - \widetilde{A}(x) - \frac{1}{2}\frac{x^2}{\tau^2} \\ &\quad - \widehat{x}S + \widetilde{A}(\widehat{x}) + \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} + \frac{(\nu+1)}{2}\log\left\{1 + \frac{(x-\widehat{x})^2}{\nu\tau_q^2}\right\} \\ &\leq \frac{1}{2}\frac{\widehat{x}^2}{\tau^2} + \{S - \widetilde{A}'(\widehat{x})\}(x - \widehat{x}) - \frac{1}{2}\frac{x^2}{\tau^2} + \frac{(\nu+1)}{2}\log\left\{1 + \frac{(x-\widehat{x})^2}{\nu\tau_q^2}\right\}.\end{aligned}$$

We recall that $\widehat{x}/\tau^2 = S - \widetilde{A}'(\widehat{x})$. Hence

$$\log \widetilde{\omega}(x, y) \leq -\frac{(x-\widehat{x})^2}{2\tau^2} + \frac{(\nu+1)}{2}\log\left\{1 + \frac{(x-\widehat{x})^2}{\nu\tau_q^2}\right\}.$$

Writing $\widetilde{x} = (x - \widehat{x})/\tau_q$ we obtain

$$\log \widetilde{\omega}(x, y) \leq -\frac{1}{2}\frac{\tau_q^2}{\tau^2}\widetilde{x}^2 + \frac{(\nu+1)}{2}\log\left(1 + \frac{\widetilde{x}^2}{\nu}\right).$$

The resulting symmetric function can be verified to be maximized at $\widetilde{x}^2 = (\nu+1)\tau^2/\tau_q^2 - \nu$, provided this expression is positive, otherwise the only maximising root is at $\widetilde{x} = 0$ and so $\log \widetilde{\omega}(x, y) \leq 0$. If the expression is positive we obtain an upper bound

$$\log \widetilde{\omega}(x, y) \leq -\frac{1}{2}\left\{(\nu+1) - \nu\frac{\tau_q^2}{\tau^2}\right\} + \frac{(\nu+1)}{2}\log\left\{\frac{(\nu+1)}{\nu}\frac{\tau^2}{\tau_q^2}\right\}.$$

COROLLARY 3. *Under the conditions of Proposition* 6 *there exists a constant* $K_3 < \infty$ *such that*

$$\sup_{\theta \in B(\overline{\theta})} E^X\{\widetilde{\omega}(X, y)^a\} \leq K_3$$

*independent of* $y$.

*Proof.* The constant in Proposition 6 depends on $\theta$ only through $\tau$. Moreover, the upper bound in Proposition 6 is continuous in $\tau$ and thus can be bounded over the compact set $B(\overline{\theta})$. □

We can summarize the results regarding the $t$-distribution in the following theorem.

THEOREM 6. *Consider the random effects model* (4.2) *and assume we have an importance sampling estimator with proposal distribution*

$$q(x \mid y) = t_\nu(x \mid \widehat{x}, \tau_q^2)$$

*with* $\tau_q^2 > 0$. *Assume additionally that either*

*i)* $\sup_x \widetilde{A}'(x) < \infty$ *or*
*ii)* $E(Y^a) < \infty$ *and* $\sup_{\theta \in B(\bar{\theta})} \widetilde{A}'(0) < \infty$.

*Then*

$$E^{Y}\left[\sup_{\theta\in B(\overline{\theta})} E^{X\mid Y}\left\{\overline{w}(Y,X,\theta)^{a}\right\}\right] < \infty.$$

*Proof.* We can bound

$$\begin{aligned} E^{Y}\left[\sup_{\theta\in B(\overline{\theta})} E^{X\mid Y}\left\{\overline{w}(Y,X,\theta)^{a}\right\}\right] &= E^{Y}\left[\sup_{\theta\in B(\overline{\theta})} \frac{E^{X\mid Y}\{\widetilde{w}(X,Y,\theta)^{a}\}}{E^{X\mid Y}\{\widetilde{w}(X,Y,\theta)\}^{a}}\right] \\ &\le E^{Y}\left[\sup_{\theta\in B(\overline{\theta})} \frac{K_3}{E^{X\mid Y}\{\widetilde{w}(X,Y,\theta)\}^{a}}\right] < \infty. \end{aligned}$$

where the first inequality is by Corollary 3 and the second by Corollary 1. □

Theorem 5 and Theorem 6 provide simple and verifiable conditions for Assumption 4 to hold in the case of generalized linear mixed models when using a Gaussian proposal or a $t$-distribution. We have established these conditions by formulating assumptions on the models and the proposal. The assumptions that are required for the model are fulfilled in the Binomial and Poisson cases as pointed out in Remark 4. Gaussian proposals require that the variance is large enough, namely

$$\tau_q^2 > \frac{1+\Delta}{2+\Delta}\tau^2,$$

where $0 < \Delta < 1$ corresponds to the quantity in Assumption 4. When one proposes from a $t$-distribution instead, no such restriction is required.

## S4. Further Simulation studies

### S4·1. *Toy example*

We consider first a simple Gaussian latent variable model where

$$X_t \sim \mathcal{N}(\theta, 1), \qquad Y_t \mid X_t = x \sim \mathcal{N}(x, 1).$$

Here $X_t, (t = 1, \ldots, T)$ are assumed to be independent. In this case, the likelihood associated to $T$ observations can be computed exactly as $p(y_{1:T} \mid \theta) = \prod_{t=1}^{T} \varphi(y_t; \theta, 2)$. This makes it an easy example to examine Assumption 1. The maximum likelihood estimator and Fisher information are given by

$$\hat{\theta}_T^{\omega} = \frac{1}{T}\sum_{t=1}^{T} Y_t, \qquad I_T(\theta) = I_T = \frac{T}{2}.$$

If we assign a zero mean Gaussian prior to $\theta$ of variance $\sigma_0^2$ then the posterior is also normal with mean $\mu_{\text{post}}$ and variance $\sigma^2_{\text{post}}$ given by

$$\mu_{\text{post}} = \left(\frac{1}{\sigma_0^2} + \frac{T}{2}\right)^{-1}\left(\frac{\sum_{t=1}^{T} Y_t}{2}\right), \qquad \sigma^2_{\text{post}} = \left(\frac{1}{\sigma_0^2} + \frac{T}{2}\right)^{-1}.$$

Assume the data are arising from the model with true parameter value $\bar{\theta}$. It follows readily from Pinsker's inequality that the Bernstein-von Mises theorem holds for $\Sigma = 2$ as we have as $T \to \infty$

$$\int \left|\pi_T^{\omega}(\theta) - \varphi\left(\theta; \hat{\theta}_T^{\omega}, I_T^{-1}\right)\right| \mathrm{d}\theta = \int \left|\varphi\left(\theta; \mu_{\text{post}}, \sigma^2_{\text{post}}\right) - \varphi\left(\theta, \hat{\theta}_T^{\omega}, \frac{2}{T}\right)\right| \mathrm{d}\theta \xrightarrow{\mathbb{P}^Y} 0.$$

Hence this model fulfils Assumption 1. To estimate the likelihood we simulate data from the model with $\bar{\theta} = 0.5$ and use $\sigma_0^2 = 10^{10}$. The likelihood is estimated using importance sampling

$$\hat{p}(y_{1:T} \mid \theta, U) = \prod_{t=1}^{T} \frac{1}{N} \sum_{i=1}^{N} \varphi(y_t - U_{t,i}; \theta, 1), \quad U_{t,i} \sim \mathcal{N}(0, 1).$$

In order to prove that Assumption 3 is fulfilled we show the stronger Assumption 4, i.e. for some $\Delta > 0$

$$E\left[\sup_{\theta \in B(\bar{\theta})} E\left\{\overline{w}(y, U, \theta)^{2+\Delta}\right\}\right] = E\left[\sup_{\theta \in B(\bar{\theta})} E\left\{\frac{\varphi(y - U; \theta, 1)^{2+\Delta}}{\varphi(y; \theta, 2)^{2+\Delta}}\right\}\right] < \infty$$

In a first step we compute for $a > 0$

$$\begin{aligned} E\left\{\frac{\varphi(y - U; \theta, 1)^a}{\varphi(y; \theta, 2)^a}\right\} &= \left(\frac{2^{a-1}}{\pi}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left(-\frac{a(y - x - \theta)^2}{2} + \frac{a(y - \theta)^2}{4} - \frac{x^2}{2}\right) \mathrm{d}x \\ &= \left(\frac{2^{a-1}}{\pi}\right)^{1/2} \int_{-\infty}^{\infty} \exp\left(-\frac{2a(y - x - \theta)^2 - a(y - \theta)^2 + 2x^2}{4}\right) \mathrm{d}x. \end{aligned}$$

Completing the square yields

$$\begin{aligned} &2a(y - x - \theta)^2 - a(y - \theta)^2 + 2x^2 \\ &= 2(a + 1)\left(x - \frac{a}{(a + 1)}(y - \theta)\right)^2 - \frac{a\,(a - 1)}{a + 1}(y - \theta)^2 \end{aligned}$$

and

$$E\left\{\frac{\varphi(y - U; \theta, 1)^a}{\varphi(y; \theta, 2)^a}\right\} = \left(\frac{2^a}{a + 1}\right)^{1/2} \exp\left\{\frac{a\,(a - 1)}{4(a + 1)}(y - \theta)^2\right\}.$$

We now consider

$$\sup_{\theta \in B(\bar{\theta})} E\left\{\overline{w}(y, U, \theta)^a\right\} = \left(\frac{2^a}{a + 1}\right)^{1/2} \exp\left\{\frac{a\,(a - 1)}{4(a + 1)} \sup_{\theta \in B(\bar{\theta})} \{(y - \theta)^2\}\right\}.$$

Now let us write

$$\begin{aligned} (y - \theta)^2 &= (y - \bar{\theta} + \bar{\theta} - \theta)^2 \\ &= (y - \bar{\theta})^2 + 2(y - \bar{\theta})(\bar{\theta} - \theta) + (\bar{\theta} - \theta)^2 \end{aligned}$$

and consider $\theta \in B(\overline{\theta})$ corresponding to $\left|\theta - \overline{\theta}\right| \leq \varepsilon$, where $\varepsilon > 0$. It is clear then that $(y - \theta)^2$ is optimised over $B(\overline{\theta})$ at either $\theta = \overline{\theta} + \varepsilon$ or $\theta = \overline{\theta} + \varepsilon$. Let us denote $y_D = y - \overline{\theta}$ and $d = \overline{\theta} - \theta$ for simplicity so that

$$(y - \theta)^2 = y_D^2 + 2y_D d + d^2,$$

Then we consider an upper bound on this which is quadratic in $y_D$ as

$$(1 + \alpha)y_D^2 + (1 + \varepsilon^2),$$

where we need to determine $\alpha$ to achieve bounding for all values $|d| \leq \varepsilon$. By symmetry of the left-hand side, we need only consider the supremum case $d = \varepsilon$ so that

$$(1 + \alpha)y_D^2 + (1 + \varepsilon^2) \geq y_D^2 + 2y_D\varepsilon + \varepsilon^2,$$

in which case, examining the roots of the resulting quadratic in $y_D$, it is required that $4\varepsilon^2 - 4\alpha \leq 0$, so $\alpha \geq \varepsilon^2$. Taking $\alpha = \varepsilon^2$ and using the bounding quadratic expression we obtain,

$$\begin{aligned}
\sup_{\theta \in B(\overline{\theta})} E[\overline{w}(y, U, \theta)^a] &= \left(\frac{2^a}{a+1}\right)^{\frac{1}{2}} \exp\left\{\frac{a(a-1)}{4(a+1)} \sup_{\theta \in B(\overline{\theta})} \{(y-\theta)^2\}\right\} \\
&\leq \left(\frac{2^a}{a+1}\right)^{\frac{1}{2}} \exp\left\{\frac{a(a-1)}{4(a+1)}(y-\overline{\theta})^2(1+\varepsilon^2) + \frac{a(a-1)}{4(a+1)}(1+\varepsilon^2)\right\} \\
&= g(y; a).
\end{aligned}$$

So finally it is required that

$$E_Y\{g(y; a)\} = \int_{-\infty}^{\infty} g(y; a)\varphi(y; \overline{\theta}, 2)\mathrm{d}y < \infty,$$

for $a = 2 + \Delta$ for some $\Delta > 0$. The above integral is finite when

$$\frac{a(a-1)}{(a+1)}(1+\varepsilon^2) < 1.$$

Hence with $a = \Delta + 2$,

$$\varepsilon^2 < \frac{(3+\Delta)}{(2+\Delta)(1+\Delta)} - 1,$$

with the right-hand side always positive provided $\Delta < \sqrt{2} - 1$.

We apply the pseudo-marginal method to this model to demonstrate how our result can approximate its characteristics. For the Markov chain, we use a random walk proposal with variance equal to the inverse Fisher information $I_T^{-1}$ scaled by $\ell = 2$. For each $T$, we run a pseudo-marginal chain for various $N$ to sample the posterior for 250000 iterations as well as the limit Markov chain of kernel $\tilde{P}_{\ell,\sigma}$. In Table S4·1 we summarize the simulations results. As expected, we find that both the average acceptance probability and the integrated autocorrelation time for $f(\theta) = \theta$ of the pseudo-marginal algorithm converge to those of the limiting Markov chain as $T$ increases.

## S4·2. *Stochastic Lotka-Volterra Model*

Assumption 3 is difficult to verify in state space models. To illustrate the applicability of our results beyond latent variable models we investigate here a stochastic kinetic Lotka-Volterra model arising in systems biology. Such models are used to describe interacting species in a predator and prey setting. In particular we consider the model with transition equations given by

$$\begin{aligned}
\mathbb{P}\left(X_{1,t+h} - X_{1,t} = 1, X_{2,t+h} - X_{2,t} = 0 \mid X_{1,t} = x_{1,t}, X_{2,t} = x_{2,t}\right) &= \beta_1 x_{1,t} + o(h) \\
\mathbb{P}\left(X_{1,t+h} - X_{1,t} = -1, X_{2,t+h} - X_{2,t} = 1 \mid X_{1,t} = x_{1,t}, X_{2,t} = x_{2,t}\right) &= \beta_2 x_{1,t} x_{2,t} + o(h) \\
\mathbb{P}\left(X_{1,t+h} - X_{1,t} = 0, X_{2,t+h} - X_{2,t} = -1 \mid X_{1,t} = x_{1,t}, X_{2,t} = x_{2,t}\right) &= \beta_3 x_{2,t} + o(h),
\end{aligned}$$

where $X_{1,t}$ and $X_{2,t}$ denotes the number of preys and predators at time $t \in [0, T]$. This model has been previously investigated, for example in (Andrieu et al., 2009) and (Wilkinson, 2012). We assume independent gamma priors for the kinetic rate parameter vector $\beta = (\beta_1, \beta_2, \beta_3)$ with

$$\beta_1 \sim \Gamma(5, 5), \quad \beta_2 \sim \Gamma(1{\cdot}5, 10), \quad \beta_3 \sim \Gamma(3{\cdot}5, 5).$$

In our simulations we assume we are only able to observe predator and prey $X_t = (X_{1,t}, X_{2,t})$ at discrete equidistant time points with independent measurement error $Y_{i,t} = X_{i,t} + W_{i,t},\ i = 1, 2,\ t = 0, \ldots, 50$ where $W_{i,t} \sim \mathcal{N}(0, 10^2)$. The artificial data have been generated using the Gillespie algorithm (Gillespie, 1977) for the rate constants $\beta = (1, 0{\cdot}005, 0{\cdot}6)$.

Table 1. *For $T$ data and $N$ particles: standard deviation $\hat{\sigma}$ of the log-likelihood estimator at $\bar{\theta}$, integrated autocorrelation time $\hat{\tau}$ and average acceptance probability $\hat{p}_{\rm acc}$ for pseudo-marginal kernel with $\ell = 2$ and limiting kernel $\tilde{P}_{\ell=2,\hat{\sigma}}$.*

| Data $T$ | Particles $N$ | $\hat{\sigma}$ | $\widehat{\text{IAT}}$ | $\hat{\text{pr}}_{\text{acc}}$ | $\widehat{\text{IAT}}\left(\tilde{P}_{\ell=2,\sigma=\hat{\sigma}}\right)$ | $\hat{\text{pr}}_{\text{acc}}\left(\tilde{P}_{\ell=2,\sigma=\hat{\sigma}}\right)$ |
|---|---|---|---|---|---|---|
| $T = 20$ | 6 | 1·70 | 17·55 | 18·69% | 31·25 | 15·32% |
| | 8 | 1·44 | 12·34 | 23·14% | 17·62 | 20·27% |
| | 10 | 1·24 | 10·76 | 26·34% | 12·44 | 24·25% |
| | 12 | 1·12 | 8·98 | 28·78% | 10·02 | 27·19% |
| $T = 30$ | 8 | 1·83 | 27·70 | 15·41% | 46·57 | 13·17% |
| | 11 | 1·47 | 16·32 | 20·24% | 18·64 | 19·61% |
| | 14 | 1·30 | 12·04 | 24·03% | 12·74 | 23·29% |
| | 17 | 1·16 | 10·85 | 26·68% | 9·91 | 26·09% |
| $T = 50$ | 20 | 1·85 | 30·46 | 13·94% | 41·53 | 13·10% |
| | 30 | 1·48 | 18·59 | 19·58% | 17·53 | 19·51% |
| | 40 | 1·29 | 13·30 | 23·59% | 11·63 | 23·34% |
| | 50 | 1·16 | 10·51 | 26·86% | 9·91 | 26·09% |
| $T = 100$ | 20 | 1·86 | 34·64 | 13·01% | 41·04 | 12·81% |
| | 30 | 1·51 | 17·98 | 19·15% | 18·73 | 18·93% |
| | 40 | 1·32 | 14·56 | 23·15% | 13·59 | 22·99% |
| | 50 | 1·16 | 10·51 | 26·33% | 9·91 | 26·09% |
| $T = 200$ | 80 | 1·83 | 38·35 | 13·11% | 46·57 | 13·17% |
| | 120 | 1·52 | 20·65 | 18·90% | 20·42 | 18·58% |
| | 160 | 1·30 | 13·87 | 22·94% | 12·74 | 23·29% |
| | 200 | 1·17 | 11·15 | 26·07% | 9·73 | 26·05% |

In this context, it is difficult to develop standard MCMC algorithms to sample the posterior distribution while the pseudo-marginal algorithm can be easily applied as an unbiased estimate of the likelihood can be computed using a bootstrap particle filter; see, e.g., (Andrieu et al., 2009) and (Wilkinson, 2012, Chapter 10). We use a multivariate Gaussian random walk proposal with scaling factor $\ell = 2{\cdot}17$ and covariance matrix close to the posterior covariance, which we estimated in a short preliminary run. This can efficiently implemented in R (R Core Team, 2017) using the package **smfsb** (Wilkinson, 2012) and the example code which can be found on the author's blog.

The algorithm is then run for 250000 iterations. We collect acceptance rate and computing time $\text{CT}(N) = \text{IAT}(N) \cdot N$ for a range of particles $N$, see Table S4·2. In practice we do not choose $\sigma\left(\bar{\theta}\right)$, but the number of particles, $N$, which is also displayed in Table S4·2. For comparison we also give an estimate of $\sigma\left(\bar{\theta}\right)$ for given $N$.

The computing time is optimized at $N = 225$ for all rates, $\beta_1$, $\beta_2$ and $\beta_3$. We estimate $\sigma\left(\bar{\theta}\right)$ to be 1·44, slightly above the results of Table 1 suggesting $\sigma = 1{\cdot}24$. The corresponding acceptance rate of 18·57% is in accordance with the one suggested by our theory, which for parameter dimension $d = 3$ yields an asymptotically optimal rate of around 19·30% ($\ell = 2{\cdot}17, \sigma = 1{\cdot}24$). We conjecture that the deviation from the results obtained in the limiting case are due to the fact that the posterior is not very concentrated around $\bar{\theta}$.

Sherlock et al. (2015) carry out Bayesian inference for a 5-dimensional stochastic Lotka-Volterra model using the pseudo-marginal algorithm based on a data set with $T = 50$ observations. The authors optimize over a grid of values for both $\sigma$ and $\ell$. Experimentally, it was found that the optimal standard deviation was $\sigma \approx 1{\cdot}45$ and the optimal tuning for the random walk achieved at $\ell = 2{\cdot}048$ with an associated optimal jumping rate of 15·39%. This is slightly above our guidelines with the values $\hat{\sigma}_{\text{opt}} = 1{\cdot}30$, $\hat{\ell}_{\text{opt}} = 2{\cdot}17$ and $\text{pr}_{\text{acc}}(\hat{\sigma}_{\text{opt}}, \hat{\ell}_{\text{opt}}) = 17{\cdot}35\%$ obtained in Table 1.

Table 2. *Comparison of the computing time for different numbers of particles in the stochastic Lotka-Volterra model.*

| Particles $N$ | Acceptance Rate | $\text{CT}(\beta_1)$ | $\text{CT}(\beta_2)$ | $\text{CT}(\beta_3)$ | $\hat{\sigma}(\bar{\theta})$ |
|---|---|---|---|---|---|
| 100 | 8·92% | 7375 | 9035 | 7564 | 2·38 |
| 125 | 11·17% | 6668 | 6717 | 6580 | 2·10 |
| 150 | 13·44% | 5805 | 5903 | 6208 | 1·84 |
| 175 | 15·62% | 5688 | 6137 | 6101 | 1·68 |
| 200 | 17·03% | 5564 | 5632 | 5744 | 1·55 |
| 225 | 18·57% | 5178 | 5452 | 5122 | 1·44 |
| 250 | 19·54% | 6107 | 6958 | 5831 | 1·36 |
| 275 | 20·82% | 5473 | 6087 | 5248 | 1·30 |
| 300 | 21·47% | 6436 | 6340 | 5959 | 1·22 |
| 325 | 22·41% | 5771 | 6586 | 6178 | 1·19 |
| 350 | 23·20% | 6406 | 6234 | 6393 | 1·13 |

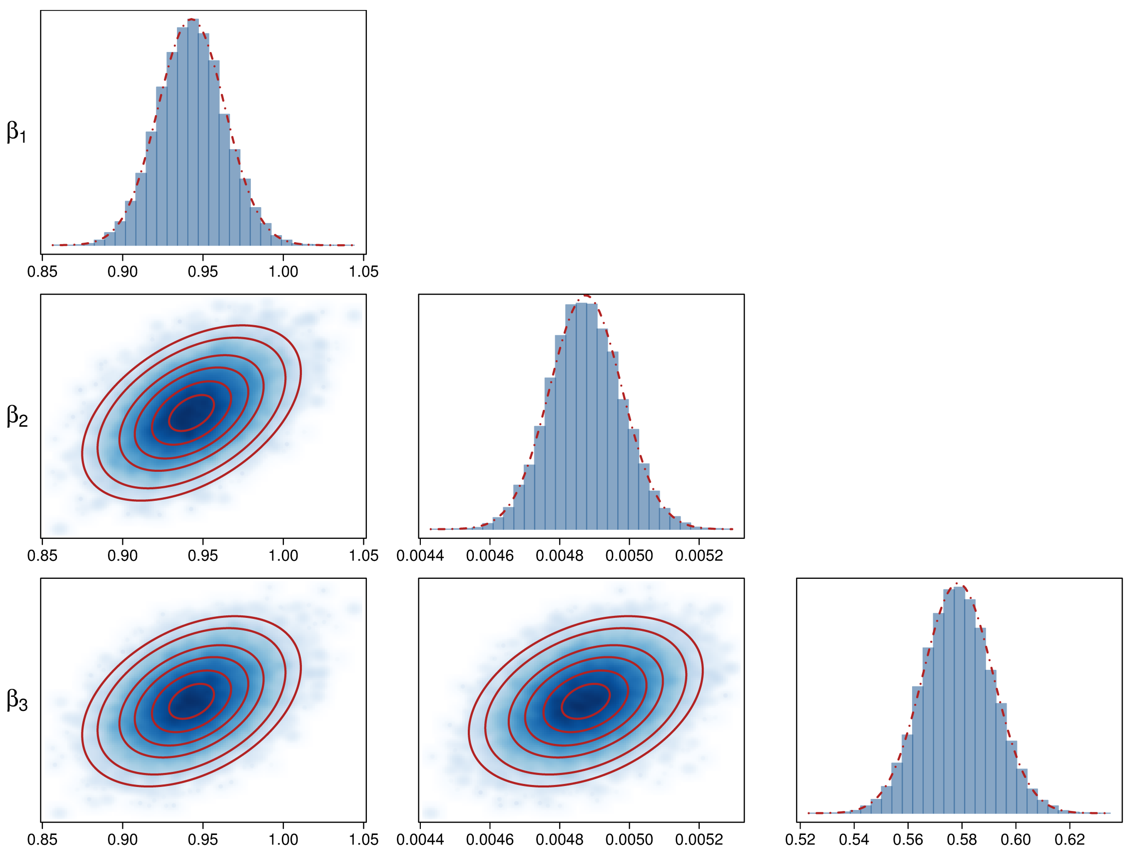


Fig. 2. Histogram of marginal posterior $p(\beta_i \mid y_{1:T})$, $i = 1, 2, 3$ on the diagonal with Gaussian approximation (line) using sample mean and variance. In addition, we show density estimates of the projections to the plane. The ellipses indicate the contour lines of a Gaussian with sample mean and sample covariance matrix. It is clear from the plots that the posterior is very close to a Gaussian.

## References


Andrieu, C., Doucet, A. & Holenstein, R. (2009). Particle Markov chain Monte Carlo for efficient numerical simulation. *Monte Carlo and quasi-Monte Carlo methods 2008* , 45–60.

Berti, P., Pratelli, L. & Rigo, P. (2006). Almost sure weak convergence of random probability measures. *Stochastics* **78**, 91–97.

Billingsley, P. (1999). *Convergence of Probability Measures*. John Wiley & Sons.

Birnbaum, Z. W. (1942). An inequality for Mill's ratio. *Ann. Math. Statist.* **13**, 245–246.

Borkar, V. S. (1991). *Topics in Controlled Markov Chains*, vol. 20. Longman Scientific & Technical UK.

Castillo, I. & Rousseau, J. (2015). Supplement to "A Bernstein–von Mises theorem for smooth functionals in semiparametric models." DOI:10.1214/ 15-AOS1336SUPP.

Chen, L. H., Goldstein, L. & Shao, Q.-M. (2010). *Normal Approximation by Stein's Method*. Springer Science & Business Media.

Crauel, H. (2003). *Random Probability Measures on Polish Spaces*. CRC Press.

Deligiannidis, G., Doucet, A. & Pitt, M. K. (2018). The correlated pseudomarginal method. *J. R. Statist. Soc.* B **80**, 839–870.

Dudley, R. M. (2002). *Real Analysis and Probability*, vol. 74. Cambridge University Press.

Ethier, S. N. & Kurtz, T. G. (2005). *Markov Processes: Characterization and Convergence*, vol. 282. John Wiley & Sons.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* **40**, 633–643.

Kallenberg, O. (2006). *Foundations of Modern Probability*. Springer-Verlag: New York.

Klenke, A. (2013). *Probability Theory: a Comprehensive Course*. Springer Science & Business Media.

Labouriau, R. (2014). A note on the identifiability of generalized linear mixed models. *arXiv preprint arXiv:1405.0673* .

McCulloch, C. E. & Neuhaus, J. M. (2005). Generalized linear mixed models. *Encyclopedia of Biostatistics* **4**.

Mukerjee, R. & Sutradhar, B. C. (2002). On the positive definiteness of the information matrix under the binary and poisson mixed models. *Ann. Instit. Statist. Math.* **54**, 355–366.

Nickl, R. (2012). Statistical Theory. *Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge* .

Owen, A. B. (2013). *Monte Carlo Theory, Methods and Examples*.

Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*, vol. 8. Cambridge University Press.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sherlock, C., Thiery, A. H., Roberts, G. O., Rosenthal, J. S. et al. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.* **43**, 238–275.

Sweeting, T. (1989). On conditional weak convergence. *J. Theoret. Probab.* **2**, 461–474.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

von Bahr, B. & Esseen, C.-G. (1965). Inequalities for the $r$th absolute moment of a sum of random variables, $1 \leqq r \leqq 2$. *Ann. Math. Statist.* **36**, 299–303.

Wilkinson, D. J. (2012). *Stochastic Modelling for Systems Biology*. CRC Press.


[*Received* 2 *January* 2017. *Editorial decision on* 1 *April* 2017]